Psychometrika

CONTENTS

VOLUME TWENTY-SIX SEPTEMBER 1961 NUMBER 3

PSYCHOMETRIKA—VOL. 26, NO. 3
SEPTEMBER, 1961

APPLICATION OF A MODEL TO PAIRED-ASSOCIATE LEARNING*

GORDON H. BOWER
STANFORD UNIVERSITY

The proposal is made to consider a paired-associate item as becoming conditioned to its correct response in all-or-none fashion, and that prior to this conditioning event the subject guesses responses at random to an unlearned item. These simple assumptions enable the derivation of an extensive number of predictions about paired-associate learning. The predictions compare very favorably with the results of an experiment discussed below.

This report describes an elementary model for the stimulus-response association process in paired-associate learning, displays an extensive number of derivations from the axioms of the model, and describes the agreement of the model with some experimental results. Paired-associate learning (PAL) as it is frequently studied involves two, at least conceptually, distinct processes: the learning of relevant responses to the general situation (e.g., as in nonsense syllable-syllable pairs), and the associative "hook-up" of these relevant responses to their appropriate stimulus members. In the belief that fractionating experimental problems leads to quicker understanding of the processes involved, this article is directed to only the second process listed above, the associative hook-up of relevant responses to their respective stimuli. The hope is that once this process is better understood the other problems, having to do with the learning of integrated response units in the situation, will become more amenable to experimental attack.

The way in which the response learning requirement was eliminated in the present experiments was to (i) use responses familiar to the subject, and (ii) inform him of the response alternatives before the experiment began. For these purposes, it was found that the first several integers (1, 2, $\cdots$, N) worked admirably. Other responses meeting the above requirements could have been used, provided precautions were taken to prevent the subject from forgetting some of the available responses during the course of the experiment. The other procedural peculiarity of these experiments was the requirement that the subject make a relevant response to each stimulus item on each trial. This procedure necessarily involved permitting the subject to control his exposure time to each stimulus.

If there are K items in the list, then a "trial" will be defined as one cycle

*This research was supported by a grant, M-3849, from the National Institutes of Mental Health, United States Public Health Service.

of presentation of each of the K items, the order of appearance of the items being randomized over successive trials. Considering only a single stimulus item for a given subject, we may categorize his responses as correct or incorrect (or, 0 or 1, as we shall use later); over successive trials there will be some particular sequence of 1's and 0's to that item. Stripped to its barest essentials, the job for a theory of PAL is to describe and account for the general characteristics of these sequences. The best job of description, of course, would be to reproduce the original sequences. Theories, as economic abstractions, do not perform this task but they can provide general descriptions (e.g., the trial number of the second success) about a sample of sequences allegedly generated under the same process laws. Obviously, models that deliver predictions about many different aspects of such sequences are preferable to less tractable models, since each prediction provides an opportunity to test the adequacy of the model. In turn, the number of predictions derivable in closed form from a model reflects to a large extent the simplicity of the assumptions used to represent the process under consideration. The assumptions of the model to be presented appear to achieve almost maximal simplicity for a model about learning; accordingly, it is possible to derive in closed form an extensive number of predictions (theorems) referring to properties of the response sequences obtained from the learning subject.

The model to be described is derived within the general framework of a stimulus sampling theory of learning [9] but with the assumption that each experimental source of stimulation (i.e., the stimulus member of a paired-associate item) may be represented by a small number of stimulus components or elements. The original investigation of small-element learning models began with a paper by Estes [10] and has been carried on by a number of people. Suppes and Atkinson [15] give an extensive development of such models and show their application to a variety of learning experiments. In the initial development of stimulus sampling theory [8, 9] it was assumed that the population of stimulus components from which the subject sampled on each trial was large. Since conditioning was assumed to take place only with respect to the sampled elements, the model implied relatively gradual changes over trials in the proportion of conditioned elements in the population and hence in response probability. Recent developments with small-element models differ in that the population of stimulus elements is assumed to be small (e.g., one or two elements) so that response probability may take on only a few values over the course of a learning experiment. The common assumption is that only one of these stimulus elements may be sampled on each trial and that the sampled element becomes conditioned to the reinforced response with probability c on every trial. Besides considerable simplification of the mathematics of stimulus sampling theory, the small-element assumptions deliver some predictions which differ markedly from

the large-element (i.e., linear) model assumptions; some of these differences are noted and will be compared with data.

The basic notion of the present model is that each stimulus item in the list of paired associates may be represented by exactly one stimulus element within the model and that the correct response to that item becomes associated in all-or-none fashion. Considering only a single item, it can be in either of two "states" on each trial: conditioned or not conditioned to the correct response. The effect of a reinforced trial (i.e., evoking the correct response in the presence of the stimulus item) is to provide an opportunity for the item to become conditioned. The single parameter of the model is c, the probability that an unconditioned item will become conditioned as the result of a reinforced trial. All items begin in the unconditioned state; the effect of continued reinforced trials is to provide repeated opportunities for the item to become conditioned.

If the item has become conditioned, then continued reinforcements of the same correct response will ensure that the item remains conditioned. The probability of the correct response when the item is conditioned is unity. The probability of the correct response when the item is not conditioned depends upon the exact experimental procedure used. In experiments by the writer, the subjects were told the N responses (integers 1, 2, $\cdots$, N) available to them and were told to respond on every trial regardless of whether they knew the correct number. If the N numbers occur equally often as the to-be-learned responses to the items, then the probability that the subject will guess correctly on an unlearned item is $1/N$; correspondingly, his probability of guessing incorrectly is $1 - (1/N)$. Our discussion of the one-element model is oriented specifically towards such an experimental procedure.

Because of the way the model is formulated, there is a partial determinism between the response sequence and the sequence of conditioning states. Specifically, if the subject responds incorrectly to a given item on trial n, then that item was not in the "conditioned" state on trial n. This feature is very helpful in deriving a number of the theorems about errors. If the subject responds correctly, however, then we cannot uniquely specify his state of conditioning, since he may have guessed correctly. Thus, it is not a consequence of the model that the subject's first correct response will be followed with probability one by correct responses on subsequent trials.

After working with the latter model for some time, it came to the writer's attention that Bush and Mosteller [6] had previously published a model for "one-trial learning" that is almost identical to the one stated above. Thus, there can be no pretense to priority in the current formulation of these elementary notions about the learning process. The present account does go beyond the abbreviated discussion by Bush and Mosteller in deriving

a large number of predictions from the model and in applying the theory with some success to verbal learning. Although their approach and the present one differ slightly in assumptions about initial conditions, the derivational techniques are sufficiently similar so that theorems can be transposed, with appropriate modifications, from one system to the other. [According to the Bush and Mosteller assumptions, a proportion c of the response sequences (subjects or items) begin in the conditioned state, and this same value of c is assumed to be the learning rate constant.]

Throughout the following sections, the predictions derived from the model will be compared with data from an experiment which now will be described. Twenty-nine subjects learned a list of ten items to a criterion of two consecutive errorless cycles. The stimuli were different pairs of consonant letters; the responses were the integers 1 and 2, each response assigned as correct to a randomly selected five stimuli for each subject. A response was obtained from the subject on each presentation of an item and he was informed of the correct answer following his response. The deck of ten stimulus cards was shuffled between trials to randomize the presentation order of the stimuli.

Axioms and Theorems about Total Errors

Axioms

1. *Each item may be represented by a single stimulus element which is sampled on every trial.*
2. *This element is in either of two conditioning states: C_1 (conditioned to the correct response) or C_0 (not conditioned).*
3. *On each reinforced trial, the probability of a transition from C_0 to C_1 is a constant, c, the probability of a transition from C_1 to C_1 is one.*
4. *If the element is in state C_1 then the probability of a correct response is one; if the element is in state C_0 , then the probability of a correct response is $1/N$, where N is the number of response alternatives.*
5. *The probability c is independent of the trial number and the outcomes of preceding trials.*

The trial to trial sequence of conditioning states forms a Markov chain, with C_1 being an absorbing state. The transition probabilities are given in the following matrix.

(1)
$$P = \begin{array}{c|cc} & C_1 & C_0 \\ \hline C_1 & 1 & 0 \\ C_0 & c & 1-c. \end{array}$$

It is easy to show that the nth power of the transition matrix is

$$(2) \qquad P^n = \begin{array}{c|cc} & C_1 & C_0 \\ \hline C_1 & 1 & 0 \\ C_0 & 1-(1-c)^n & (1-c)^n. \end{array}$$

We explicitly assume that all items start out in state C_0 (i.e., are not conditioned initially). Thus, starting out in state C_0, the probability of still being in state C_0 after n reinforced trials is $(1-c)^n$, which approaches zero as n becomes large. Thus, for $c > 0$, with probability one the process will eventually end in conditioning state C_1 (i.e., will become conditioned).

For each item, define a sequence of response random variables, x_n, which take on the value 1 if an error occurs on trial n, or the value 0 if a success occurs on n. From the axioms, the conditional probabilities of an error given states C_1 or C_0 at the beginning of trial n are

$$(3) \qquad \Pr\{x_n = 1 \mid C_{1,n}\} = 0 \quad \text{and} \quad \Pr\{x_n = 1 \mid C_{0,n}\} = 1 - \frac{1}{N}.$$

To obtain the average probability of an error on the nth trial, q_n, multiply these conditional probabilities by the probabilities of being in C_1 or C_0, respectively, at the start of trial n:

$$(4) \qquad \begin{aligned} q_n = \Pr\{x_n = 1\} &= \Pr\{x_n = 1 \mid C_{1,n}\} \Pr\{C_{1,n}\} \\ &\quad + \Pr\{x_n = 1 \mid C_{0,n}\} \Pr\{C_{0,n}\} \\ &= 0 + \left(1 - \frac{1}{N}\right)(1-c)^{n-1} = \left(1 - \frac{1}{N}\right)(1-c)^{n-1}. \end{aligned}$$

The expected total number of errors, u_1, before perfect learning is given by

$$(5) \qquad u_1 = E\left[\sum_{n=1}^{\infty} x_n\right] = \sum_{n=1}^{\infty} \Pr\{x_n = 1\} = \sum_{1}^{\infty} \left(1 - \frac{1}{N}\right)(1-c)^{n-1} = \frac{1 - \frac{1}{N}}{c}.$$

The expected total errors per item serves as a stable estimator of c. For the experiment described above with $N = 2$, the average number of errors per item was 1.45. Equating u_1 in (5) to 1.45, the c value obtained is .344. This estimate of c will be fixed throughout the remaining discussion of these data. Using this value of c in (4), the predicted learning curve in Fig. 1 is obtained.

In the expression for u_1, all errors are weighted equally. It is also possible to derive expressions for various weighted sums of errors, as Bush and Sternberg [7] have shown for the linear model. The results here are identical with

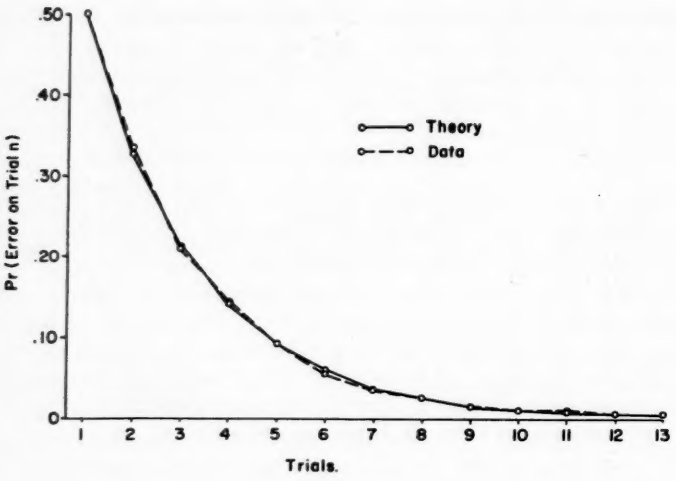


FIGURE 1

q_n, the Probability of an Incorrect Response over Successive Trials of the Experiment

their results. Three examples of the expectation of weighted error sums are given below.

$$(6) \qquad E\left[\sum_{n=1}^{\infty} nx_n\right] = \sum_{n=1}^{\infty} n\left(1 - \frac{1}{N}\right)(1 - c)^{n-1} = \frac{1 - \frac{1}{N}}{c^2} = \frac{u_1}{c};$$

$$(7) \qquad E\left[\sum_{n=1}^{\infty} \frac{x_n}{n}\right] = \sum_{n=1}^{\infty} \frac{\left(1 - \frac{1}{N}\right)(1 - c)^{n-1}}{n} = \frac{1 - \frac{1}{N}}{1 - c} \log \frac{1}{c};$$

$$(8) \qquad E\left[\sum_{n=1}^{\infty} \frac{x_n}{(n-1)!}\right] = \left(1 - \frac{1}{N}\right) \sum_{m=0}^{\infty} \frac{(1 - c)^m}{m!} = \left(1 - \frac{1}{N}\right) e^{1-c}.$$

It is possible to obtain the distribution of the total number of errors on each item. This distribution was derived by Bush and Mosteller; their result is readily translated into the terms of the current approach to the theory. If we let T represent the total number of errors (to perfect learning) on a single item, the probability distribution of T is

$$(9) \qquad \Pr\{T = k\} = \begin{cases} b/N & \text{for } k = 0 \\ \dfrac{b(1 - b)^k}{1 - c} & \text{for } k \geq 1, \end{cases}$$

where

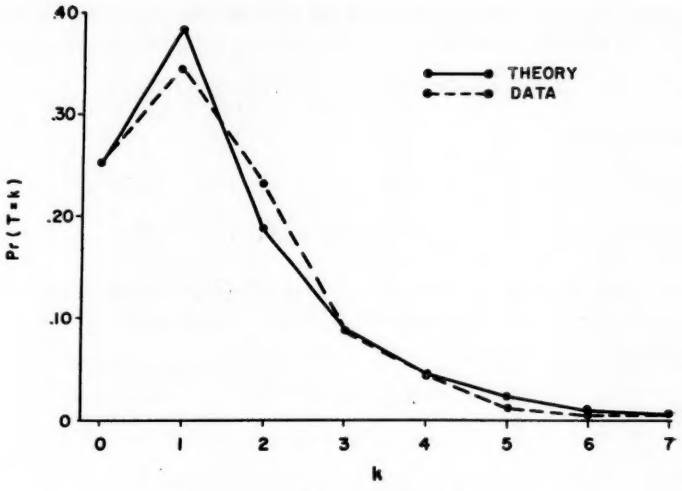


FIGURE 2
Distribution of T, the Total Number of Errors per Item

$$b = \frac{c}{1 - \frac{1 - c}{N}}.$$

The mean of T was derived as u_1 in (5); the variance of T is given by

$$\text{Var}(T) = u_1 + (1 - 2c)u_1^2 . \tag{10}$$

The predicted and obtained distributions of T are shown in Fig. 2.

Sequential Properties of the Model

Predictions about sequential features of the data may be obtained by considering runs of errors. To date only mean values of the various run distributions have been derived; higher moments will not be discussed. Let r_j represent the number of error runs of length j in an infinite number of trials; we seek the expectation of r_j . For these purposes, it is convenient to define another random variable, u_j , which counts the number of j-tuples of errors that occur in an infinite sequence of trials. Formally, define u_j as

$$u_j = \sum_{n=1}^{\infty} x_n x_{n+1} \cdots x_{n+j-1} \quad \text{for} \quad j = 1, 2, \cdots . \tag{11}$$

The product, $x_n x_{n+1} \cdots x_{n+j-1}$, has the value one only when j consecutive errors occur starting with the error on trial n. It may be seen that u_1 is just the total number of errors.

To make clear how the u_j are being counted and their relation to the r_j, consider the possible sequence

$$1111100110001101000 \cdots \text{ (all the rest zeros).}$$

For this sequence,

$$u_1 = 10, \quad u_2 = 6, \quad u_3 = 3, \quad u_4 = 2, \quad u_5 = 1;$$
$$r_1 = 1, \quad r_2 = 2, \quad r_3 = r_4 = 0, \quad r_5 = 1, \quad R = \sum_j r_j = 4.$$

R is the total number of error runs. In an excellent article, Bush [5] has shown that the r_j can be expressed as linear combinations of the u_j. In particular,

$$r_j = u_j - 2u_{j+1} + u_{j+2}, \tag{12}$$

and

$$R = \sum_{j=1}^{\infty} r_j = u_1 - u_2. \tag{13}$$

Having expressed the r_j in terms of the u_j, we now turn to deriving from the model the expected value of u_j. We proceed as follows:

$$E(u_j) = E\left[\sum_{n=1}^{\infty} x_n \cdot x_{n+1} \cdots \cdot x_{n+j-1}\right] = \sum_{n=1}^{\infty} \Pr\{x_n = 1\} \tag{14}$$
$$\cdot \Pr\{x_{n+1} = 1 \mid x_n = 1\} \Pr\{x_{n+2} = 1 \mid x_n \cdot x_{n+1} = 1\} \cdots$$
$$\cdot \Pr\{x_{n+j-1} = 1 \mid x_n \cdot x_{n+1} \cdots x_{n+j-2} = 1\}.$$

Because of the Markovian properties of the model, the lengthy conditional probabilities on the right-hand side can be simplified, viz.,

$$\Pr\{x_{n+i} = 1 \mid x_n = 1, x_{n+1} = 1, \cdots, x_{n+i-1} = 1\} \tag{15}$$
$$= \Pr\{x_{n+i} = 1 \mid x_{n+i-1} = 1\}.$$

That is, if the subject made an error on the preceding trial, then that is all the information there is to be extracted from the entire preceding sequence of responses. His error tells us that his conditioning state on the preceding trial was C_0; the probability of an error on the current trial is then

$$\Pr\{x_{n+1} = 1 \mid x_n = 1\} \tag{16}$$
$$= c \cdot 0 + (1 - c)\left(1 - \frac{1}{N}\right) = (1 - c)\left(1 - \frac{1}{N}\right) = \alpha,$$

and, moreover, this holds for any trial number n. Thus, using relations (15) and (16), the equation for u_j becomes

$$E(u_j) = \sum_{n=1}^{\infty} \Pr\{x_n = 1\} \Pr\{x_{n+1} = 1 \mid x_n = 1\} \cdots \cdot \Pr\{x_{n+j-1} = 1 \mid x_{n+j-2} = 1\} \tag{17}$$

$$= \sum_{n=1}^{\infty} \Pr\{x_n = 1\} \underbrace{\alpha \cdots \alpha}_{(j-1) \text{ times}}.$$

$$E(u_j) = \alpha^{j-1} \sum_{n=1}^{\infty} \Pr\{x_n = 1\} = u_1\alpha^{j-1}.$$

With these values in hand, now calculate R and r_j, using relations (12) and (13).

$$E(R) = E(u_1) - E(u_2) = u_1(1 - \alpha), \tag{18}$$

$$E(r_j) = E(u_j) - 2E(u_{j+1}) + E(u_{j+2}) = u_1(1 - \alpha)^2\alpha^{j-1} \tag{19}$$

$$= R(1 - \alpha)\alpha^{j-1}.$$

Another useful summary of sequential properties in the data is the extent to which an error on trial n tends to be followed by an error k trials later, without regard to what responses intervene between trials n and $n + k$. Define $c_{k,n}$ as $x_n \cdot x_{n+k}$; this expression will have the value 1 only if errors occur on both trials n and $n + k$. It may be noted that $c_{k,n}$ summarizes the same features as does an autocorrelation of lag k. The expectation of $c_{k,n}$ is

$$E(c_{k,n}) = E(x_n \cdot x_{n+k}) = E(x_{n+k} \mid x_n) \cdot E(x_n) \tag{20}$$

$$= \Pr\{x_{n+k} = 1 \mid x_n = 1\} \Pr\{x_n = 1\}.$$

To find the conditional probability above, note that for an error to occur on trial $n + k$ it must be the case that conditioning has failed to occur during the intervening k trials, and moreover that the subject guesses incorrectly on trial $n + k$. The probability of this joint event is

$$\Pr\{x_{n+k} = 1 \mid x_n = 1\} = (1 - c)^k\left(1 - \frac{1}{N}\right). \tag{21}$$

Therefore,

$$E(c_{k,n}) = \left(1 - \frac{1}{N}\right)(1 - c)^k\left(1 - \frac{1}{N}\right)(1 - c)^{n-1}. \tag{22}$$

A convenient statistic for comparison with data is obtained by taking the

"autocorrelation" of x_n and x_{n+k} over all trials n of the experiment. Defining c_k as the mean value of this random variable,

$$(23) \qquad c_k = E\left[\sum_{n=1}^{\infty} x_n x_{n+k}\right] = \sum_{n=1}^{\infty} E(c_{k,n}) = u_1\left(1 - \frac{1}{N}\right)(1 - c)^k$$

$$\text{for} \quad k = 1, 2, 3, \cdots .$$

Predicted and observed values of c_1 , c_2 , and c_3 are given in Table 1.

It is a simple matter to construct other statistics which capture various features of the sequential dependencies in the response sequence. Such statistics are expressible as various sums and/or products of the x_n . One illustration will be provided here to demonstrate the general derivational techniques. In order to predict the average number of alternations of successes and failures that occur over the response sequence, define a random variable A_n which will count an alternation between trials n and $n + 1$. Hence,

$$(24) \qquad A_n = (1 - x_n)x_{n+1} + x_n(1 - x_{n+1}).$$

It will be noted that A_n takes on the value 1 either if a success occurs on trial n and a failure on trial $n + 1$ or if a failure occurs on n and a success on $n + 1$. Multiplying out and taking the expectation of A_n yields

$$(25) \qquad E(A_n) = \frac{\alpha}{N}(1 - c)^{n-1} + (1 - \alpha)\left(1 - \frac{1}{N}\right)(1 - c)^{n-1}.$$

The average of the sum of A_n over trials is

$$(26) \qquad A = E[\sum A_n] = u_1\left[c + \frac{2(1 - c)}{N}\right].$$

Errors during Various Parts of Learning

In this section we derive the distribution of the number of errors between the kth and $(k + 1)$st success and also of the number of errors between the kth and $(k + 2)$nd success. As special cases of these general results, for $k = 0$ we obtain the distributions of errors before the first and before the second success. The methods employed in these derivations are general so that the distribution of errors between the kth and $(k + m)$th success could be obtained, the sole limitation being that the expressions get progressively more cumbersome as m is increased.

Consider first the distribution of the number of errors occurring between the kth and $(k + 1)$st success. Let J_k be this random variable; it can take on the values 0, 1, 2, $\cdots$ of the non-negative integers. Errors following the kth success can occur only if the kth success itself came about by guessing (rather than via prior conditioning). Thus, the probability that the kth success occurred by guessing (call it g_k) will play a central role in the expres-

sion for the distribution of J_k. To forego for the moment the derivation of g_k, write the distribution of J_k as

$$\Pr\{J_k = i\} = \begin{cases} 1 - g_k\alpha & \text{for } i = 0 \\ g_k(1-\alpha)\alpha^i & \text{for } i > 0. \end{cases} \tag{27}$$

For example, the probability of three errors between the kth success and the next one is given by the joint probability of (i) the kth success occurred by guessing, (ii) conditioning failed to occur at the end of trials, k, $k+1$, and $k+2$ and incorrect guesses occurred on trials $k+1$, $k+2$, and $k+3$, the probability of this joint event being $(1-c)^3\,(1-1/N)^3 = \alpha^3$, and (iii) given that the element was not conditioned at the start of trial $k+3$, a correct response occurs on trial $k+4$ with probability $1-\alpha$. To obtain the term for $J_k = 0$, note that no errors could occur either if the kth success occurred via prior conditioning (with probability $1 - g_k$) or, having guessed the kth success, a success occurs on the next trial with probability $1-\alpha$. The sum of these two terms, $1 - g_k$ and $g_k(1-\alpha)$, gives the probability that $J_k = 0$.

From the distribution in (27) one obtains the mean and variance of J_k.

$$E(J_k) = \frac{\alpha g_k}{1-\alpha}, \qquad \text{Var}(J_k) = \frac{\alpha g_k}{(1-\alpha)^2}[1 + \alpha(1 - g_k)]. \tag{28}$$

The task now is to derive g_k, the probability that the kth success occurs by guessing. Consider g_1, the probability that the first success occurs by guessing. It is

$$\begin{aligned} g_1 &= \frac{1}{N} + \left(1 - \frac{1}{N}\right)(1-c)\frac{1}{N} + \left(1 - \frac{1}{N}\right)^2(1-c)^2\frac{1}{N} + \cdots \\ &= \frac{1}{N}\sum_{j=0}^{\infty}\alpha^j = \frac{1}{N(1-\alpha)}. \end{aligned} \tag{29}$$

That is, the subject guesses correctly on the first trial with probability $1/N$; he may fail there so the item does not become conditioned and he guesses correctly on the second trial, and so on. It can be shown for $k > 1$ that a general recursion holds for g_k, viz.,

$$g_k = g_{k-1}(1-c)\left[\frac{1}{N} + \alpha\frac{1}{N} + \alpha^2\frac{1}{N} + \cdots\right] = g_{k-1}(1-c)g_1. \tag{30}$$

That is, for the kth success to occur by guessing, it must be the case that (i) the $(k-1)$st success occurred by guessing, (ii) conditioning failed to occur on the trial of the $(k-1)$st success, and (iii) starting out not conditioned on the next trial, the next correct response also occurs by guessing, with probability g_1.

Equation (30) is a standard linear difference equation having the solution

$$g_k = (1 - c)^{k-1} g_1^k = g_1 \left(1 - \frac{c}{1 - \alpha}\right)^{k-1}. \tag{31}$$

Since $c < 1 - \alpha$, it follows that g_k decreases exponentially with k. This result is intuitively clear: the tenth success is less likely to occur by guessing than is, say, the second success. Corresponding to the decrease in g_k, the average errors between the kth and $(k + 1)$st success is decreasing exponentially over k, as (28) shows.

We have been considering J_k for $k > 0$. The interpretation of J_0 is the number of errors before the first success. It is convenient to define g_0 as $1/(1 - c)$, although g_0 itself has no physical interpretation. Defining g_0 in this way, then the distribution of J_0, the errors before the first success, is given by (27). The distribution of J_0 has more intuitive appeal when written as

$$\Pr\{J_0 = i\} = \begin{cases} \dfrac{1}{N} & \text{for } i = 0 \\ \left(1 - \dfrac{1}{N}\right)(1 - \alpha)\alpha^{i-1} & \text{for } i \geq 1, \end{cases} \tag{32}$$

although formally it is the same as (27) with $g_0 = 1/(1 - c)$.

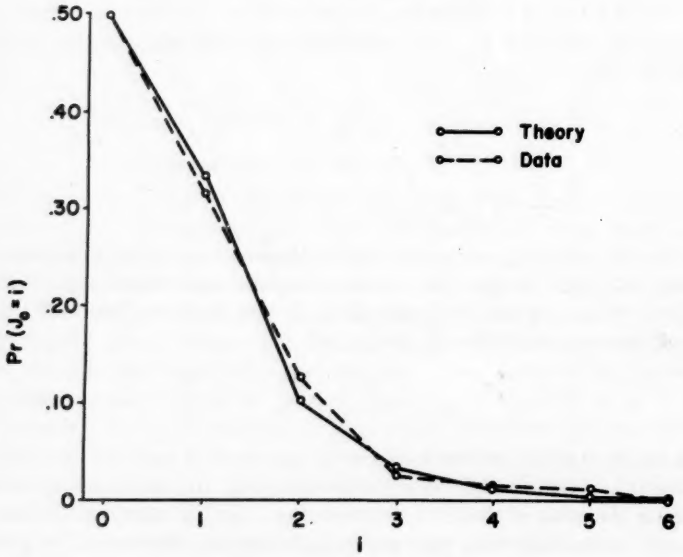


FIGURE 3
Distribution of J_0, the Number of Errors Before the First Success

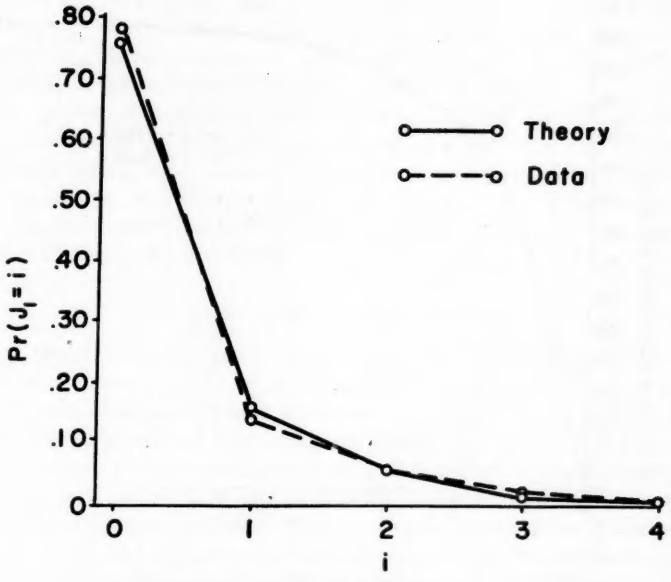


FIGURE 4

Distribution of J_1, the Number of Errors Between the First and Second Success

To illustrate the fit of the model to data, the distribution of the number of errors before the first success is shown in Fig. 3, and the mean and standard error, predicted and observed, are shown in Table 1. Also, the theoretical and observed distributions of J_1, the number of errors between the first and second success, are shown in Fig. 4.

Using the J_k values so calculated, one obtains an expression for the average errors before the kth success. If F_k is defined as the cumulative errors before the kth success, then the obvious recursion on the means is

$$E(F_{k+1}) = E(F_k) + E(J_k). \tag{33}$$

The solution of this difference equation is

$$E(F_k) = \sum_{i=0}^{k-1} E(J_i) = \frac{\alpha}{1-\alpha} \sum_{i=0}^{k-1} g_i \,. \tag{34}$$

Substituting the values for g_i, the summation yields

$$\begin{aligned} E(F_k) &= \frac{1-\frac{1}{N}}{1-\alpha} + \frac{\alpha g_1}{1-\alpha} \frac{[1-(g_1(1-c))^{k-1}]}{[1-g_1(1-c)]} \\ &= u_1 - \frac{\alpha}{Nc(1-\alpha)} [g_1(1-c)]^{k-1}, \end{aligned} \tag{35}$$

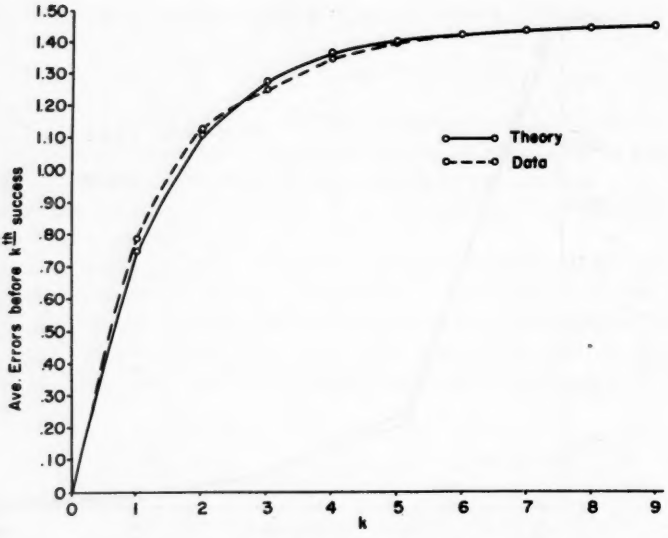


FIGURE 5
$E(F_k)$, the Expected Number of Errors Before the kth Success

where u_1 is the average total errors per item as given in (5). Equation (35) establishes the expected result that, for large k, the average number of errors before the kth success approaches the average total number of errors per item. In Fig. 5, the observed and predicted values of $E(F_k)$ through the ninth success are shown.

The distribution of the number of errors between the kth and $(k + 2)$nd success has been obtained and is presented here for completeness. Define S_k as the number of errors between the kth and $(k + 2)$nd success; it is clear that $S_k = J_k + J_{k+1}$. By specialization for $k = 0$, S_0 gives the distribution of the number of errors before the second success. The distribution of S_k, which is given here without proof (see [3]), is

$$\text{(36)} \qquad \Pr\{S_k = i\} = \begin{cases} 1 - g_k + g_k\left[c + \dfrac{(1-c)(1-\alpha)}{N}\right] & \text{for } i = 0 \\ g_k\left[c + \dfrac{(1-c)(1-\alpha)}{N}(i+1)\right]\alpha^i & \text{for } i \geq 1, \end{cases}$$

and the first and second raw moments of the distribution are

$$\text{(37)} \qquad E(S_k) = \frac{\alpha g_k}{(1-\alpha)^2}\left[c + \frac{2(1-c)}{N}\right],$$

$$E(S_k^2) = \frac{\alpha g_k}{(1-\alpha)^3}[2(1-\alpha)(1+2\alpha) - c(1+3\alpha)].$$

The g_k are as given before. Again, defining $g_0 = 1/(1 - c)$, (36) gives the distribution of the number of errors before the second success. The observed and predicted distributions of S_0 are shown in Fig. 6, and the mean and standard error, predicted and observed, are given in Table 1.

The preceding derivations have been carried out for the number of errors before the kth success, etc. The number of *trials* before the kth success is obviously related by a constant. Thus, the trial number of the kth success is the number of errors before the kth success, F_k , plus k. Changing to a "trial" notation shifts the origin (adds a constant) but does not affect the form or variance of the distribution.

The Trial Number of the Last Failure

Our purpose in this section is to derive the distribution of the trial number of the last error in an effectively infinite sequence of trials. However, before proceeding with this derivation, it is helpful to consider another statistic: the proportion of items characterized by having no errors following the first success. In the experimental data, a considerable percentage (62.8 percent, in fact) of the item protocols displayed this characteristic and the question arose whether the model would predict such results. Let p_1 represent the probability that a response sequence will exhibit this property of no errors following the first success. If b represents the probability of no more errors following a correct guess, then an expression for p_1 is

$$p_1 = 1 - g_1 + g_1 b = 1 - g_1(1 - b). \tag{38}$$

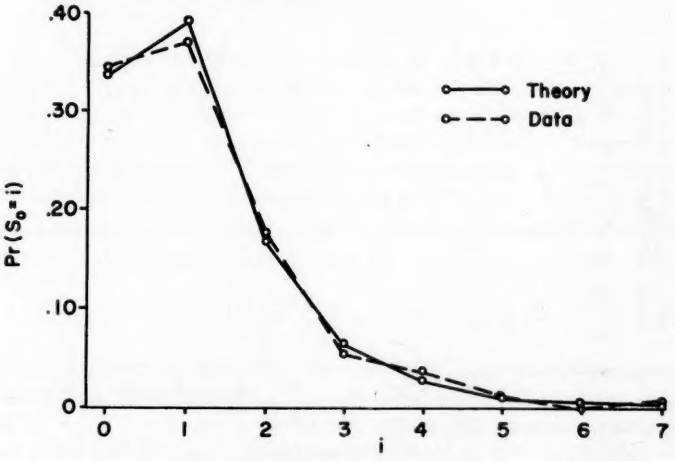


FIGURE 6
Distribution of S_0, the Number of Errors Before the Second Success

That is, a proportion $1 - g_1$ of the first correct responses come about via prior conditioning (so no more errors will occur), while $g_1 b$ represents the probability that the first correct response occurs by guessing but no more errors occur. To complete this derivation, b, the probability that no errors occur following a correct guess is

$$b = c + (1 - c)\frac{1}{N}c + (1 - c)^2\left(\frac{1}{N}\right)^2 c + \cdots \tag{39}$$

$$= \frac{c}{1 - \frac{(1 - c)}{N}} = \frac{c}{\alpha + c}.$$

That is, with probability c the item was conditioned on the trial on which the correct guess occurred; with probability $1 - c$ conditioning failed to occur on that trial, the subject guessed correctly on the next trial with probability $1/N$ and the item became conditioned then with probability c, and so on. This value of b is the same as that used in the distribution of the number of errors given in (9).

Substituting this result for b into (38),

$$p_1 = 1 - \frac{g_1\alpha}{\alpha + c}. \tag{40}$$

Using the estimate of c obtained earlier, the predicted p_1 is .638, which is quite close to the observed proportion of .628.

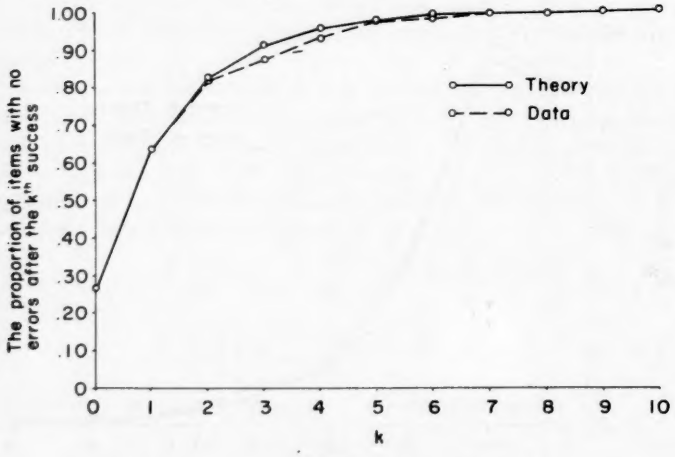


FIGURE 7

p_k, the Probability of Zero Errors Following the kth Success

As (40) suggests, define p_k to be the probability that there are no errors following the kth success. Using our previous result for g_k, one derives

$$p_k = 1 - \frac{\alpha g_k}{\alpha + c} = 1 - \frac{\alpha g_1}{\alpha + c}[g_1(1 - c)]^{k-1}. \tag{41}$$

Observed and predicted values of p_k are shown in Fig. 7.

To determine the position of the last error, define n' as the random variable representing the trial number on which the last error occurs in an infinite sequence of trials. If no errors occur at all, then n' is set equal to zero. The probability distribution of n' is

$$\Pr\{n' = k\} = \begin{cases} \dfrac{b}{N} & \text{for } k = 0 \\ b\left(1 - \dfrac{1}{N}\right)(1 - c)^{k-1} & \text{for } k \geq 1. \end{cases} \tag{42}$$

The first value is just $\Pr\{T = 0\}$, which was given in (9). If some errors occur, then for the last error to occur on trial k it must be the case that conditioning failed to occur on the preceding $k - 1$ trials, an incorrect guess occurred on trial k, but no errors followed that, with probability b. The mean and variance of n' are

$$E(n') = m = \frac{b\left(1 - \dfrac{1}{N}\right)}{c^2} = \frac{bu_1}{c}, \tag{43}$$

$$\text{Var}(n') = m\left[\frac{2}{c} - 1 - m\right].$$

The observed and predicted distributions of n' are shown in Fig. 8; the mean and standard error, observed and predicted, are given in Table 1.

Consider now the distribution of the number of successes that intervene between the kth and $(k + 1)$st error, provided that a $(k + 1)$st error occurs. Because an error effectively "resets" the process to state C_0, the distribution of this random variable will be independent of k and of the trial number on which the leading error occurs. Let H represent the number of intervening successes. The distribution of H is given by

$$\Pr\{H = j\} = \frac{\alpha}{1 - b}\left[\frac{1 - c}{N}\right]^j = (\alpha + c)(1 - \alpha - c)^j, \; j = 0, 1, 2, \cdots. \tag{44}$$

The division by $1 - b$ establishes the condition that at least one more error will occur. The probability that the next error occurs on the very next trial is just $(1 - c)(1 - 1/N) = \alpha$; the probability that the next response is a correct guess and the error occurs on the following trial is $(1/N)(1 - c)\alpha$, and so on. Although the derivation of the number of successes before the

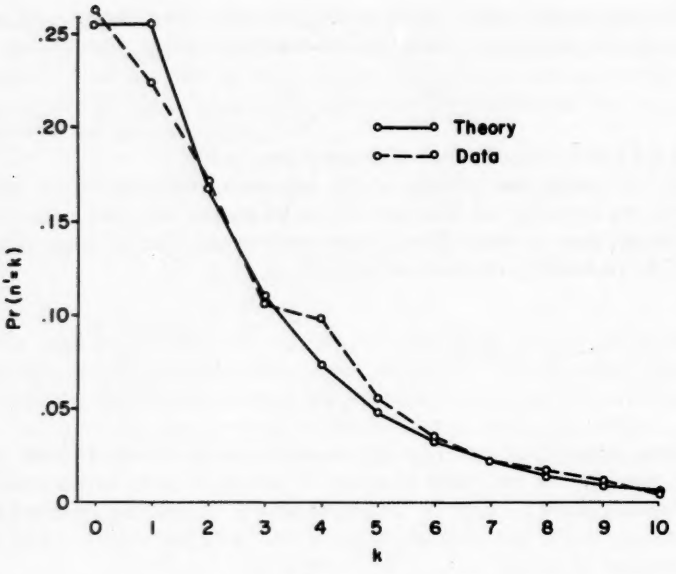


FIGURE 8
Distribution of n', the Trial Number of the Last Failure

first error (provided there is one) proceeds somewhat differently, the resulting distribution is identical to the distribution given in (44). The mean and variance of H are

$$E(H) = \frac{1 - \alpha - c}{\alpha + c}, \qquad \text{Var}(H) = \frac{1 - \alpha - c}{(\alpha + c)^2}. \tag{45}$$

The observed and predicted distributions of H are shown in Fig. 9, and the means and standard errors are given in Table 1.

The preceding analyses have been carried out for the responses to a single item over trials. If the items can be considered homogeneous in difficulty so that each learning process may be characterized by the same c value, then it is possible to derive a number of predictions about performance across items within a particular trial. If there are K items, then a run through the list (a trial) provides a sequentially ordered sample of size K from a binomial population. With this characterization, one can then derive various quantities of experimental interest, e.g., the probability that a perfect recitation of the list occurs on trial n, the average number of error runs of length j considering the K sequential samples on each trial, and so on. Derivations of such results are relatively easy and are presented in [3].

Comparison with Linear Model

What has been accomplished in the preceding sections is a detailed analysis of the sequence of response random variables. In terms of sheer bulk of predictions derivable from learning axioms, the sole comparable alternative is the single-operator linear model explored extensively by Bush and Sternberg [7]. It would be instructive, therefore, to place on record a detailed quantitative comparison of the fit of these two models to the present data. The basic notion of the linear model is that the associative strength between a stimulus and its correct response increases by a linear transformation following each reinforced trial. Stated differently, the probability of an error is expected to decrease by a constant fraction following each reinforced trial. If the initial error probability is $1 - (1/N)$ then over successive reinforced trials the error probability decreases, taking on a number of values intermediate between $1 - (1/N)$ and 0. In contrast, the one-element model proposed here assumes that the error probability has only two values, $1 - (1/N)$ or 0, and jumps from the first to the second value following the trial on which the all-or-none association is formed.

Although these conceptions differ markedly, the two models predict the same average learning curve. Thus, finer details of the data are required to differentiate these models. Since, according to the linear model, q_n decreases by the same fraction every trial, the response random variables, x_n ,

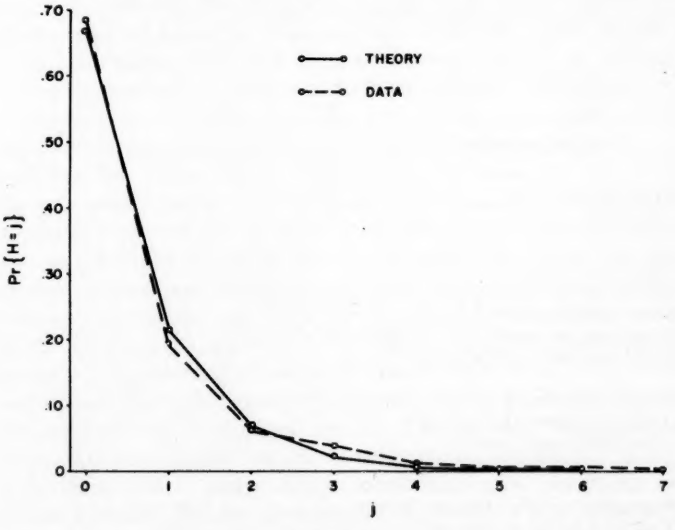


FIGURE 9

Distribution of H, the Number of Successes Intervening Between Adjacent Errors

are statistically independent; that is, the probability of an error on trial n is expected to be the same whether the subject responded correctly or incorrectly on the preceding trial. For the one-element model the x_n are not independent; whether we expect an error on trial n depends heavily on whether or not an error occurred on the preceding trial. Noting these differences, we are led to expect that the two models would be differentiated best by their predictions about sequential aspects of the data. Indeed this is the case, as may be seen in Table 1 which collects 19 comparisons of the one-element and linear models with data. The linear model predictions were obtained by referring to the theorems derived by Bush and Sternberg [7]. Three other

TABLE 1

Comparison of One-Element and Linear Models with Data

	Statistic	One element	Data	Linear
1.	Ave. errors per item	---	1.45	---
2.	S. D.	1.44	1.37	1.00
3.	Ave. errors before first success	.749	.785	.705
4.	S. D.	.98	1.08	.84
5.	Ave. errors between first and second success	.361	.350	.315
6.	S. D.	.76	.72	---
7.	Ave. errors before second success	1.11	1.13	1.02
8.	S. D.	1.10	1.01	.93
9.	Ave. successes between errors	.488	.540	---
10.	S. D.	.72	.83	---
11.	Ave. trial of last error	2.18	2.33	3.08
12.	S. D.	2.40	2.47	3.39
13.	Total error runs	.975	.966	1.162
14.	Error runs of length 1	.655	.645	.949
15.	Error runs of length 2	.215	.221	.144
16.	Error runs of length 3	.070	.058	.064
17.	Error runs of length 4	.023	.024	.005
	Autocorrelation of errors			
18.	--one trial apart (c_1)	.479	.486	.288
19.	--two trials apart (c_2)	.310	.292	.195
20.	--three trials apart (c_3)	.201	.187	.127
21.	Alternations of success and failure	1.45	1.43	1.83
22.	(Fail-Fail runs) – (Success-Fail runs)	.000	.020	-.380
23.	Probability of a success following an error	.672	.666	.730

statistics are shown for which the predictions of the linear model have not been worked out (although stat rats could have been run for these).

The results in Table 1 require little comment. Of the 19 possible comparisons between the one-element and linear models, the one-element model comes closer to the data on 17. The greatest differentiation of the models is seen in sequential statistics, lines 13 through 23, and in the trial number of the last failure (lines 11 and 12). The largest absolute discrepancy from data of the one-element predictions occurs with the average trial number of the last failure, but this statistic also has the largest variance of all those considered. Weighing these considerations along with the excellent fits of the one-element model to the data shown in Figs. 2–9, we may conclude that the one-element model provides a more adequate description of these data than does the linear model.

Other paired-associate data favoring the one-element model have been reported in [4]. One dramatic comparison of the two models is provided by considering the expected number of errors (to perfect learning) following an error that occurs on trial n. According to the linear model, the number of errors expected following an error on trial n should be a decreasing function of n, since associative strength is assumed to increase steadily with the number of preceding reinforced trials. In contrast, from the one-element model the expectation is that the average errors following an error on trial n is a constant, $(1 - c)u_1$, which is independent of the trial number on which the error was observed. The point of the matter is that if we observe an error on trial n, then we know the item was not conditioned prior to that trial; hence, we can assume that our learning process "starts" in conditioning state C_0 at the beginning of trial n and that the state of the subject's associative connection has not effectively changed since he started the experiment. We may, so to speak, reset the clock back to the beginning of the experiment for predicting the subject's future behavior on that item.

To get a stable test of these different predictions, the present data from 29 subjects were pooled with the data of 47 other subjects learning 10 paired-associate items under the same conditions except for 14 of the subjects the number of response alternatives was 3, and for 14 there were 8 responses. The varying N's would not affect the constancy or monotone decreasing aspects of the two predictions. For the 760 learning sequences the average number of errors following an error on trial 1, on trial 2, $\cdots$, on trial 6 were calculated. The data beyond trial 6 were not analyzed since the number of cases involved was dropping off rapidly. The results of these calculations are shown in Fig. 10 where the one-element model prediction (i.e., average of all the data points) and a rough approximation to the linear model's predictions are included for comparative purposes. There is little doubt that the one-element prediction is closer to the data, which show remarkable constancy.

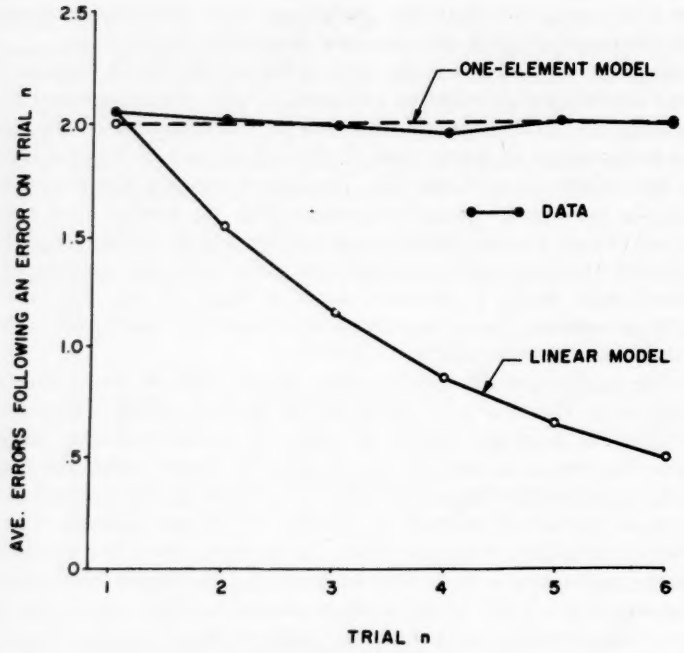


FIGURE 10
Average Number of Errors Following an Error on Trial n
(The one-element and linear predictions are indicated.)

The predicted function for the linear model is not exact since groups with differing N's and c's were pooled in Fig. 10; however, the function does show the relative order of magnitude of differences to be expected from the linear model. The values for the graph were obtained by estimating the average c value across groups (it was .25) and then multiplying successive values on the function by $1 - c$. For example, the average errors observed following an error on trial 1 was 2.05; hence, for trial 2 the linear prediction would be 2.05 (.75) = 1.54, and so on.

Goodness of Fit Considerations

Although the preceding tabulation of various statistics and distributions tells us something about how well the model describes these data, still one legitimately may raise the question of whether there is some summary measure for evaluating the *over-all* goodness of fit of the model to these data. For these purposes a chi-square procedure adapted for stochastic learning

models by Suppes and Atkinson [15] from an original paper by Anderson and Goodman [1] may be used. This procedure tests the ability of a model to reproduce the nth order conditional probabilities in the response sequences. Tests involving such quantities have priority in evaluating goodness of fit since the primary reference of stochastic models is to these conditional probabilities. Other statistics of the data (error runs, trial of second success, etc.) are more or less derived from these conditional probabilities and thus should have less priority in deciding over-all goodness of fit. The chi-square test proposed by Anderson and Goodman is most appropriate for those models which assume that the sequence of response random variables is a finite Markov chain (i.e., that current response probability depends upon, say, only one or two prior responses). This happens to be a rather restricted class of learning models; however, the test is practically useful even for chains of infinite order in which current response probability presumably depends upon the entire past history of responses and reinforcements. In practice, such chains can be approximated reasonably well by taking account of only a small number (say 3 or 4) of prior trials when calculating conditional probabilities from the theory.

The chi-square procedure may be illustrated with the present data. The decision was made to look at third-order conditional probabilities using the data from the first six trials of the experiment; beyond trial 6 practically all responses were correct so very little information could be gained by considering the data beyond that point. With two responses (correct and incorrect) there are eight possible sequences of length three. The data were tabulated in an 8×2 table, the entries in each cell corresponding to the frequency with which a given sequence of responses on trials n, $n + 1$, and $n + 2$ was followed by a success (or failure) on trial $n + 3$. For each subject-item sequence, three observations were obtained corresponding to n taking on the values 1, 2, and 3. There were thus $3(290) = 870$ observations in total.

The expected conditional probabilities are readily calculated from the one-element model. For example, four of the eight prior sequences have an error on trial $n + 2$; hence, the expected conditional probability of an error on trial $n + 3$ would be α. The only conditional probability which is troublesome to compute is that of an error given a prior sequence of three successes (with responses prior to trial n being unspecified). This conditional probability is calculated separately for $n = 1, 2, 3$, and then the three results are averaged. Analogous computations from the linear model are extremely simple—in that model the x_n are considered to be statistically independent; hence, one merely averages the response probabilities on trials 4, 5, and 6.

The conditional probabilities calculated above are converted into cell frequencies by multiplying them by the observed frequency of a given prior sequence of three responses (i.e., we multiply by the observed row sums of the table). Chi-square values can then be calculated separately for the one-

element model and for the linear model. There are eight rows in the table, each row having one linear constraint (the two entries must sum to the appropriate row total) and for each model we have estimated one parameter (c); hence, each chi square will have seven degrees of freedom. The chi-square values for the observed and expected frequencies were 9.40 for the one-element model and 98.36 for the linear model. Therefore, the test rejects the linear model in its fit to these data but does not reject the one-element model.

Suppes and Chmura [16] have proposed a simple but rigorous procedure for discriminating between the goodness of fit of two models for which the above chi-square values have been calculated. Their statistic, T, is the ratio of the two chi-square values, each divided by its respective degrees of freedom. Under the assumption that one of the models is true, T is distributed as the noncentral F statistic, with a noncentrality parameter equal to the value of an ordinary chi square done on the two sets of expected frequencies (ignoring the data for the moment). For the present case, the value of T is 10.40. This value is so large (an ordinary F table requires only 7.00 for significance at the .01 level) that it would be a mere formality to calculate its exact probability under the assumption that both models fit the data equally well. Hence, we may unequivocally reject the linear model in favor of the one-element model.

Range of Application of the Model

The fact that the one-element model gives an adequate quantitative account of these paired-associate data satisfies one important requisite of a scientific theory, that of being close to the data. If, in addition, the theory is mathematically tractable in that numerous consequences are easily derived in closed form, then indeed we are in a fortunate position. The main task of this paper has been to show that the one-element model is mathematically tractable; those familiar with current work in mathematical learning theory certainly can have no quarrel with this claim. This property of the model is due to the extreme simplicity of its assumptions about the association process. One might effectively argue that the present model nearly achieves the absolute minimum in assumptions for a workable theory of learning.

Once one has demonstrated the predictive validity of a model for a limited class of experimental situations, there remains the task of characterizing more generally those experimental arrangements to which the model may be expected to apply. In the first part of this report, we explicitly restricted the model to the S-R association process and have used simplified experimental situations in which response learning was precluded. Within this restricted domain of PAL, the model has proved extremely useful in investigating the effects on learning of variations in the number of response alternatives and in the reinforcement conditions prevailing during learning [4].

In addition, the model has led us to do experiments in which the guessing probabilities are altered indirectly by varying the proportion of items in the list that have the same correct response (e.g., with 20 items and responses 1 and 2, we have varied the number of items that have 1 as the correct response).

The experimental conditions may differ considerably from those obtaining under paired-associate learning, but still the model may be expected to apply if response learning is precluded. A good example of such an application is to the paradigm that experimenters have called verbal discrimination learning (e.g., Rundquist and Freeman, [14]). In one variant of this experiment, the subject is required to read the correct response from a card on which are printed N alternatives (words, syllables); the subject goes repeatedly through a deck of K such cards until he can give the correct response to all of them. The model has been applied to the results of such an experiment with $N = 2$; its predictive validity proved equally as good as that reported here for the paired-associate task. To cite a further example of work in progress, we are attempting to extend the model to a similar task in which the subject learns to recognize or identify a visual form as one of those that had been shown to him in a "training list" of visual forms.

A further extension of the present work would investigate the modifications in the theory that are required to handle those PAL situations in which the responses per se must be learned. Here again it may prove advantageous to fractionate the problem by utilizing experimental arrangements which primarily involve only response learning. The free verbal recall paradigm [e.g., 2] would appear to serve these purposes. In such experiments the subject is read a number of unrelated words and later is tested for free, unaided recall. With this arrangement, the responses are conditioned presumably to situational and intraverbal cues in a manner analogous to that assumed to occur in PAL response learning. Evidence already exists to indicate that the free verbal recall situation may yield to a simple theoretical analysis. Miller and McGill [12] and Murdock [13] have published quantitative theories which appear to account adequately for their results from free verbal recall experiments. Ultimately, one would like to have a set of combination axioms whereby the assumptions about S-R association and response learning may be combined for predicting results in those experimental situations involving the concurrent operation of these two processes. It may not be presumptuous to suppose that such a development will come about in the next few years.

REFERENCES

[1] Anderson, T. W. and Goodman, L. A. Statistical inference about Markov chains. *Ann. math. Statist.*, 1957, **28**, 89-110.

[2] Bruner, J. S., Miller, G. A., and Zimmerman, C. Discriminative skill and discriminative matching in perceptual recognition. *J. exp. Psychol.*, 1955, **49**, 187-192.

[3] Bower, G. H. Properties of the one-element model as applied to paired-associate

learning. Tech. Rep. No. 31, Psychol. Ser., Inst. for Mathematical Studies in the Social Sciences, Stanford Univ., 1960.

[4] Bower, G. H. A model for response and training variables in paired-associate learning. *Psychol. Rev.*, in press.

[5] Bush, R. R. Sequential properties of linear models. In R. R. Bush and W. K. Estes (Eds.), *Studies in mathematical learning theory*. Stanford: Stanford Univ. Press, 1959. Pp. 215-227.

[6] Bush, R. R. and Mosteller, F. A comparison of eight models. In R. R. Bush and W. K. Estes (Eds.), *Studies in mathematical learning theory*. Stanford: Stanford Univ. Press, 1959. Pp. 293-307.

[7] Bush, R. R. and Sternberg, S. A single-operator model. In R. R. Bush and W. K. Estes (Eds.), *Studies in mathematical learning theory*. Stanford: Stanford Univ. Press, 1959. Pp. 204-214.

[8] Estes, W. K. Toward a statistical theory of learning. *Psychol. Rev.*, 1950, **57**, 94-107.

[9] Estes, W. K. and Burke, C. J. A theory of stimulus variability in learning. *Psychol. Rev.*, 1953, **60**, 276-286.

[10] Estes, W. K. Component and pattern models with Markovian interpretations. In R. R. Bush and W. K. Estes (Eds.), *Studies in mathematical learning theory*. Stanford: Stanford Univ. Press, 1959. Pp. 9-52.

[11] Estes, W. K. Learning theory and the new mental chemistry. *Psychol. Rev.*, 1960, **67**, 207-223.

[12] Miller, G. A. and McGill, W. J. A statistical description of verbal learning. *Psychometrika*, 1952, **17**, 369-396.

[13] Murdock, B. B. The immediate retention of unrelated words. *J. exp. Psychol.*, 1960, **60**, 222-234.

[14] Rundquist, W. N. and Freeman, M. Roles of association value and syllable familiarization in verbal discrimination learning. *J. exp. Psychol.*, 1960, **59**, 396-401.

[15] Suppes, P. and Atkinson, R. C. *Markov learning models for multi-person interactions*. Stanford: Stanford Univ. Press, 1960.

[16] Suppes, P. and Chmura, H. A statistical test for comparative goodness of fit of alternative learning models. Tech. Rep. No. 36, Psychol. Ser., Inst. for Mathematical Studies in the Social Sciences, Stanford Univ., 1961.

Manuscript received 9/1/60

Revised manuscript received 1/11/61

PSYCHOMETRIKA—VOL. 26, NO. 3
SEPTEMBER, 1961

A GENERALIZATION OF STIMULUS SAMPLING THEORY

RICHARD C. ATKINSON
STANFORD UNIVERSITY

A modification of stimulus sampling theory is presented. The restriction that each stimulus element is conditioned to one and only one response is replaced with the notion of a scale of conditioning for each element. This variation provides a context in which such variables as reward magnitude and motivation can be viewed as determiners of behavior. Some experimental results on multiple response problems also have a natural interpretation in terms of these ideas.

The phrase "Stimulus Sampling Theory" is used to describe various formulations of the basic theory first set forth by Estes [2] and Estes and Burke [3]. To date, all models that have been derived from this theory are characterized by the assumption that the stimuli (components or patterns) are conditioned in an all-or-none fashion to responses. There have been no theoretical or empirical developments which clearly indicate that an all-or-none postulate is inadequate to account for learning phenomena; nevertheless it is of scientific interest to examine alternative formulations where the assumption of all-or-none conditioning is replaced by a strength-of-conditioning process.

The purpose of this paper is to examine a natural generalization of stimulus sampling notions which incorporates a conditioning strength variable. We shall introduce such a variable with reference to a particular set of axioms derived from stimulus sampling theory, namely, the axioms given by Suppes and Atkinson ([5], see ch. 1). Their Conditioning Axioms C4 and C5 and Sampling Axioms S1 and S2 are to remain unchanged; only axioms C1, C2, C3, and R1 are to be modified. These modifications generate rather surprising predictions and provide a new context within which such variables as reward magnitude and motivation can be analyzed. Further, some experimental results on multiple response problems have a natural interpretation in terms of the ideas presented in this paper.

We begin by stating the axioms for the two-response case since it is the simplest; the generalization to multiple responses will be examined later. As customary, the responses are denoted A_1 and A_2, and three reinforcing events E_0, E_1, and E_2 are specified. The first group of axioms deals with the conditioning of stimuli, the second group with the sampling of stimuli, and the third with responses.

Conditioning Axioms

C1. *Associated with each stimulus element i is a positive integer s_i . (s_i denotes the maximum value of conditioning strength.)*

C2. *At the start of trial n, stimulus element i is in conditioning state K_j , where $j = 0, 1, 2, \cdots, s_i$.*

C3. *If stimulus element i is sampled on trial n and is in conditioning state K_j , then with probability $1 - \theta$ the reinforcing event is not effective and no change occurs in the conditioning state. When the reinforcing event is effective (i.e., with probability θ) then the conditioning state*

(a) *changes to K_{j+1} if E_1 occurs (however, if in K_{s_i} on trial n then no change occurs),*

(b) *changes to K_{j-1} if E_2 occurs (however, if in K_0 on trial n then no change occurs),*

(c) *remains unchanged if E_0 occurs.*

C4. *Stimulus elements which are not sampled on a trial do not change their conditioning state on that trial.*

C5. *The probability θ is independent of the trial number and the preceding pattern of events.*

Sampling Axioms

S1. *Exactly one stimulus element is sampled on each trial.*

S2. *Given the set of elements available for sampling on a trial, the probability of sampling a particular element is independent of the trial number and the preceding pattern of events.*

Response Axiom

R1. *If stimulus element i is in conditioning state K_j and the element is sampled, then the probability of an A_1 response is j/s_i .*

These axioms are formally identical to those given by Suppes and Atkinson [5] when $s_i = 1$ for all elements. For this special case, methods of estimating the number of elements (N) and the conditioning parameter θ have been worked out; many applications to empirical data are available. When $s_i > 1$ for some elements, then interesting and rather surprising predictions occur. We now proceed to examine this case. In much of the discussion we shall restrict ourselves to the one-element model ($N = 1$). There are no mathematical problems in extending the analysis to the multi-element case but notation becomes extremely complex. Further, a consideration of the one-element case is adequate for illustrating the basic ideas.

Noncontingent Reinforcement

We begin with the simple noncontingent situation where E_0's are not permitted and the probability of events E_1 and E_2 are constant over trials;

i.e., $P(E_{1,n}) = \pi \geq \frac{1}{2}$. We may prove from our axioms that the sequence of random variables which take the conditioning states as values is a Markov chain. This means, among other things, that a transition matrix $P = [p_{ij}]$ may be constructed, where $p_{ij} = P(K_{j,n+1} \mid K_{i,n})$, and $K_{x,n}$ denotes the event of being in state K_x on trial n. The learning process is completely characterized by these transition probabilities and the initial probability distribution on the conditioning states.

By Axiom C3, it is obvious that

$$
\begin{aligned}
&p_{s,s} = 1 - \theta + \theta\pi, \qquad p_{s,s-1} = \theta(1 - \pi), \\
&\left.\begin{aligned} p_{i,i+1} &= \theta\pi \\ p_{i,i} &= 1 - \theta \\ p_{i,i-1} &= \theta(1 - \pi) \end{aligned}\right\} \quad \text{for} \quad i \neq 0, s, \qquad (1) \\
&p_{0,1} = \theta\pi, \qquad p_{0,0} = 1 - \theta + \theta(1 - \pi).
\end{aligned}
$$

(For the one-element case we supress the subscript in the notation s_1 and simply write s.)

Next define $p_{ij}^{(n)}$ as the probability of being in state j on trial $n + 1$, given that on trial 1 the process was in state i. Moreover, if the appropriate limit exists and is independent of i, we set

$$
u_j = \lim_{n \to \infty} p_{ij}^{(n)}. \qquad (2)
$$

The Markov chain defined by (1) is irreducible and aperiodic; for such a finite-state chain it is well known that the limiting quantities u_j exist. For our particular case

$$
u_j = \begin{cases} \dfrac{a^{s-j} - a^{s-j+1}}{1 - a^{s+1}} & \text{for} \quad a \neq 1 \\ \dfrac{1}{s + 1} & \text{for} \quad a = 1, \end{cases} \qquad (3)
$$

where $a = (1 - \pi)/\pi$.

By the Response Axiom R1, the asymptotic probability of an A_1 response in the noncontingent situation is

$$
\begin{aligned}
\lim_{n \to \infty} P(A_{1,n}) = P(A_1) &= \sum_{j=0}^{s} \binom{j}{s} u_j \qquad (4) \\
&= \frac{s(1 - a) - a(1 - a^s)}{s(1 - a)(1 - a^{s+1})} \quad \text{for} \quad \pi \neq \tfrac{1}{2} \\
&= \tfrac{1}{2} \qquad \text{for} \quad \pi = \tfrac{1}{2}.
\end{aligned}
$$

For $\pi = \frac{1}{2}$ the prediction of $P(A_1)$ is $\frac{1}{2}$ for all values of s. However for $\pi \neq \frac{1}{2}$ the asymptotic prediction depends on s. Fig. 1 presents $P(A_1)$ as a function of π; the parameter on each curve is the value of s. For s equal to 1,

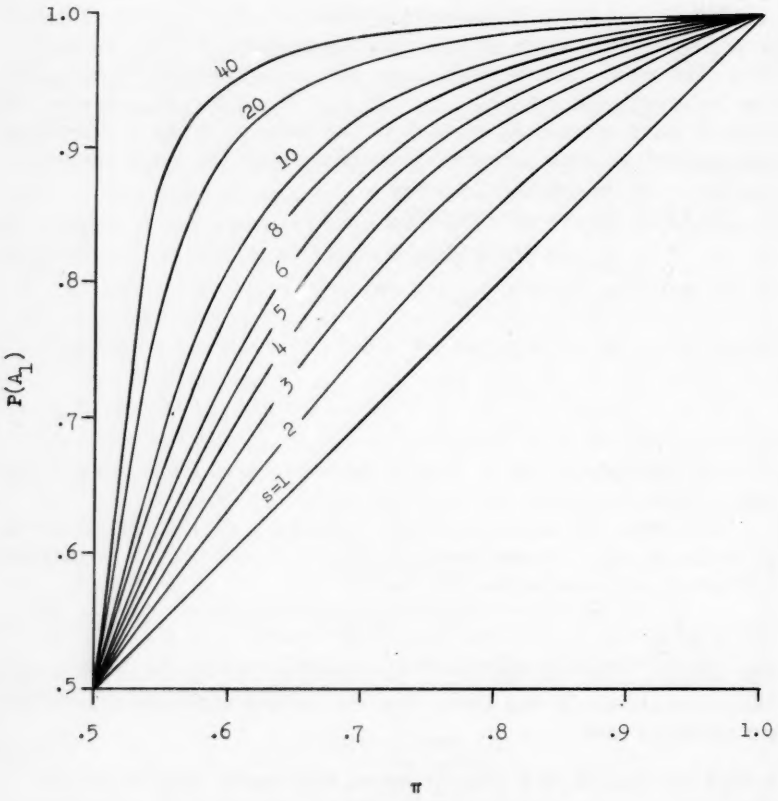


FIGURE 1

Predicted Asymptotic Probability of an A_1 Response as a Function of π

$P(A_1) = \pi$; however as s increases, the prediction for $P(A_1)$ becomes increasingly greater than π. In fact by inspection of (4) it is obvious that

$$\lim_{s\to\infty} P(A_1) = 1 \quad \text{for} \quad \pi > \tfrac{1}{2}.$$

Comparable results can be obtained for other reinforcement schedules. For example, consider a contingent situation where E_0's are not permitted and let $P(E_{1,n} \mid A_{1,n}) = \pi_1$ and $P(E_{1,n} \mid A_{2,n}) = \pi_2$. For this case if

$$\frac{\pi_2}{1 - \pi_1 + \pi_2} > \tfrac{1}{2},$$

then $P(A_1)$ approaches 1 as s becomes large. For example, if $\pi_1 = \frac{3}{4}$ and $\pi_2 = \frac{1}{2}$, then $P(A_1)$ is .67, .71, .75, .79, $\cdots$ for $s = 1, 2, 3, 4, \cdots$.

Suppes and Atkinson ([5], see ch. 10) report data for a noncontingent experiment where $\pi = .6$. The independent variable was the amount of money won or lost on each trial when the subject was correct ($A_{1,n}E_{1,n}$ or $A_{2,n}E_{2,n}$) or incorrect ($A_{2,n}E_{1,n}$ or $A_{1,n}E_{2,n}$). For subjects in Group Z, no money was won or lost; for Group F five cents was won when the subject was correct and the same amount lost when incorrect; for Group T ten cents was won or lost. The obtained proportions of A_1 responses at asymptote (trials 141–240) were .593 (Group Z), .644 (Group F), and .690 (Group T). If we were to estimate s for the one-element model from these data alone, we would find that s is approximately 1.0 for Group Z, 2.3 for Group F, and 3.3 for Group T. Thus, for this experiment the estimated value of s increased as a function of the monetary payoff. In terms of the elementary process the amount of change in response probability on a given trial is dependent on the monetary payoff. For example, in the one-element model if $P(A_{1,n}) = 0$, an E_1 occurs, and conditioning is effective then $P(A_{1,n+1}) = 1/s$. Of course, these ideas apply directly to experimental situations where different amounts of money can be won or lost from trial to trial; more detailed notions concerning the relations of θ and s to monetary value will depend on this type of experiment.

These asymptotic results for the one-element model can be generalized to the multi-element case and thereby permit $P(A_1)$ to take any value in the interval from π to 1. Specifically, if there are N elements, then

$$P(A_1) = \begin{cases} \frac{1}{N}\sum_{i=1}^{N} \frac{s_i(1-a) - a(1-a^{s_i})}{s_i(1-a)(1-a^{s_i+1})} & \text{for } \pi \neq \tfrac{1}{2} \\ \tfrac{1}{2} & \text{for } \pi = \tfrac{1}{2}. \end{cases} \tag{5}$$

It should be noted that for $N > 1$ and any set of values for s_i ($i = 1, \cdots, N$) we have a chain of infinite order in the sequence of response random variables; the same statement holds for $N = 1$ and $s > 1$. However, for the special case where $N = s = 1$, the sequence of response random variables is a first-order Markov chain.

We shall not pursue the multi-element case but instead turn to some sequential results for the one-element noncontingent model. We present only a few to illustrate the method of proof and have selected those quantities which are useful in making estimates of θ and s. (See [5, ch. 2] for a discussion of appropriate estimation procedures.)

Consider first $P(A_{1,n+1} \mid E_{1,n}A_{1,n})$. By elementary probability considerations and Axiom R1,

$$\begin{aligned} &P(A_{1,n+1}E_{1,n}A_{1,n}) \\ &= \sum_{i,j} P(A_{1,n+1}K_{j,n+1}E_{1,n}A_{1,n}K_{i,n}) \\ &= \sum_{i,j} P(A_{1,n+1} \mid K_{j,n+1})P(K_{j,n+1} \mid E_{1,n}A_{1,n}K_{i,n})P(E_{1,n})P(A_{1,n} \mid K_{i,n})P(K_{i,n}). \end{aligned}$$

However, by Axiom C3,

$$P(A_{1,n+1}E_{1,n}A_{1,n}) = \sum_{i=0}^{s-1}\left[\frac{i+1}{s}\theta + \frac{i}{s}(1-\theta)\right]\pi\frac{i}{s}P(K_{i,n}) + \pi P(K_{s,n})$$

$$= \frac{\theta\pi}{s}\sum_{i=0}^{s-1}\frac{i}{s}P(K_{i,n}) + \pi\sum_{i=0}^{s-1}\frac{i^2}{s^2}P(K_{i,n}) + \pi P(K_{s,n}).$$

Note, however, that

$$\lim_{n\to\infty} P(K_{i,n}) = u_i$$

and by (3)

$$\lim_{n\to\infty} P(A_{1,n+1}E_{1,n}A_{1,n}) = \frac{\theta\pi}{s}[P(A_1) - u_s] + \pi[V_2 - u_s] + \pi u_s$$

$$= \frac{\theta\pi}{s}[P(A_1) - u_s] + \pi V_2 ,$$

where

$$V_2 = \sum_{i=0}^{s}\left(\frac{i}{s}\right)^2 u_i$$

and can be easily calculated. Thus

$$\lim_{n\to\infty} P(A_{1,n+1} \mid E_{1,n}A_{1,n}) = \frac{1}{P(A_1)}\left\{\frac{\theta}{s}[P(A_1) - u_s] + V_2\right\}. \tag{6}$$

Other asymptotic predictions useful for estimating parameters may be obtained by similar arguments.

$$\lim_{n\to\infty} P(A_{1,n+1} \mid E_{2,n}A_{2,n}) = \frac{1}{P(A_2)}\left\{P(A_1) - V_2 + \frac{\theta}{s}[u_0 - P(A_2)]\right\},$$

$$\lim_{n\to\infty} P(A_{1,n+1} \mid E_{2,n}A_{1,n}) = \frac{V_2}{P(A_1)} - \frac{\theta}{s}, \tag{7}$$

$$\lim_{n\to\infty} P(A_{1,n+1} \mid E_{1,n}A_{2,n}) = \frac{1}{P(A_2)}\left\{\frac{\theta}{s}P(A_2) + P(A_1) - V_2\right\}.$$

The derivation of mean learning curves and other sequential statistics for this model are presented in [1]. It should be noted, however, that the mean learning curve may or may not be exponential in form, the particular shape of the function will depend on initial-condition parameters and the s_i values.

Multiple Responses

We now examine the case where there are r responses (A_1 , $\cdots$, A_r) and $r + 1$ reinforcing events (E_0 , E_1 , $\cdots$, E_r). For the multiple-response case it is necessary to restate axioms C2, C3, and R1 more generally.

C2′. *At the start of trial n, stimulus element i is in conditioning state* $\langle k_{1,n} k_{2,n} \cdots k_{r,n} \rangle$, *where* $k_{i,n} = 0, 1, \cdots, s_i$ *and* $k_{1,n} + k_{2,n} + \cdots + k_{r,n} = s_i$.

C3′. *If stimulus element i is sampled on trial n and is in conditioning state* $\langle k_{1,n} \cdots k_{r,n} \rangle$, *then with probability* $1 - \theta$ *the reinforcing event is not effective and no change occurs in the conditioning state. When the reinforcing event is effective (i.e., with probability* θ*)*

(a) *if* $E_{x,n}$ $(x \neq 0)$ *occurs, then* $k_{x,n+1} = k_{x,n} + 1$ *and one and only one of the other k's takes a decrement of* 1. *The probability (for* $j \neq x$*) that* $k_{j,n+1} = k_{j,n} - 1$ *is* $k_{j,n}/(s_i - k_{x,n})$,

(b) *if* $E_{0,n}$ *occurs, then the conditioning state remains unchanged (i.e.,* $k_{i,n+1} = k_{i,n}$).

R1′. *If stimulus element i is in conditioning state* $\langle k_{1,n} \cdots k_{r,n} \rangle$ *and the element is sampled, then the probability of response* A_j *is* $k_{j,n}/s_i$.

For $r = 2$ these axioms are equivalent to the axioms given at the outset of this paper. The only reason for introducing the earlier version was to make the presentation of the two-response case more accessible.

We now apply the axioms to a noncontingent reinforcement procedure reported by Gardner [4]. Three responses (A_1, A_2, A_3) are available to the subject and three reinforcing events (E_1, E_2, E_3) are employed. On each trial one of the reinforcing events occurs, i.e., $P(E_{i,n}) = \pi_i$, where $\pi_1 + \pi_2 + \pi_3 = 1$. Again, we consider only the one-element case, but there are no mathematical problems in extending this analysis to multiple elements; the only difficulty is that notation and computations can become very involved.

First consider the case where $s = 1$. There are three conditioning states $\langle 100 \rangle$, $\langle 010 \rangle$, and $\langle 001 \rangle$. These states form a Markov chain whose transition matrix can be obtained from Axiom C3′. This matrix is

$$\begin{array}{c} \\ \langle 100 \rangle \\ \langle 010 \rangle \\ \langle 001 \rangle \end{array} \begin{array}{c} \begin{array}{ccc} \langle 100 \rangle & \langle 010 \rangle & \langle 001 \rangle \end{array} \\ \begin{bmatrix} 1 - \theta + \theta\pi_1 & \theta\pi_2 & \theta\pi_3 \\ \theta\pi_1 & 1 - \theta + \theta\pi_2 & \theta\pi_3 \\ \theta\pi_1 & \theta\pi_2 & 1 - \theta + \theta\pi_3 \end{bmatrix} \end{array}.$$

Define u_{ijk} $(i, j, k = 1, 0)$ analogous to (2). Then by Axiom R1′

$$\begin{aligned} \lim_{n \to \infty} P(A_{1,n}) &= P(A_1) = u_{100} = \pi_1 , \\ \lim_{n \to \infty} P(A_{2,n}) &= P(A_2) = u_{010} = \pi_2 , \\ \lim_{n \to \infty} P(A_{3,n}) &= P(A_3) = u_{001} = \pi_3 . \end{aligned} \tag{8}$$

For $s = 2$, the conditioning states are $\langle 200\rangle$, $\langle 110\rangle$, $\langle 101\rangle$, $\langle 020\rangle$, $\langle 011\rangle$, and $\langle 002\rangle$; the transition matrix is

$$
\begin{array}{c|cccccc}
 & \langle 200\rangle & \langle 110\rangle & \langle 101\rangle & \langle 020\rangle & \langle 011\rangle & \langle 002\rangle \\
\langle 200\rangle & 1-\theta+\theta\pi_1 & \theta\pi_2 & \theta\pi_3 & & & \\
\langle 110\rangle & \theta\pi_1 & 1-\theta & \tfrac{1}{2}\theta\pi_3 & \theta\pi_2 & \tfrac{1}{2}\theta\pi_3 & \\
\langle 101\rangle & \theta\pi_1 & \tfrac{1}{2}\theta\pi_2 & 1-\theta & & \tfrac{1}{2}\theta\pi_2 & \theta\pi_3 \\
\langle 020\rangle & & \theta\pi_1 & & 1-\theta+\theta\pi_2 & \theta\pi_3 & \\
\langle 011\rangle & & \tfrac{1}{2}\theta\pi_1 & \tfrac{1}{2}\theta\pi_1 & \theta\pi_2 & 1-\theta & \theta\pi_3 \\
\langle 002\rangle & & & \theta\pi_1 & & \theta\pi_2 & 1-\theta+\theta\pi_3
\end{array}.
$$

It can be shown that

$$
\begin{aligned}
u_{200} &= \pi_1^2/A, & u_{020} &= \pi_2^2/A, \\
u_{110} &= \pi_1\pi_2/A, & u_{011} &= \pi_2\pi_3/A, \\
u_{101} &= \pi_1\pi_3/A, & u_{002} &= \pi_3^2/A,
\end{aligned}
$$

where $A = \pi_1^2 + \pi_2^2 + \pi_3^2 + \pi_1\pi_2 + \pi_1\pi_3 + \pi_2\pi_3$. By Axiom R1′

$$
\begin{aligned}
P(A_1) &= u_{200} + \tfrac{1}{2}[u_{110} + u_{101}] = \pi_1[\pi_1 + \tfrac{1}{2}(1-\pi_1)]/A, \\
P(A_2) &= u_{020} + \tfrac{1}{2}[u_{110} + u_{011}] = \pi_2[\pi_2 + \tfrac{1}{2}(1-\pi_2)]/A, \\
P(A_3) &= u_{002} + \tfrac{1}{2}[u_{101} + u_{011}] = \pi_3[\pi_3 + \tfrac{1}{2}(1-\pi_3)]/A.
\end{aligned} \tag{9}
$$

The analysis may be extended to any value of s (see [1]). For r responses the number of conditioning states will be $\binom{r+s-1}{s}$. However, for our examination of the Gardner data a comparison of predictions for s equal to 1 and 2 will be sufficient.

Gardner actually reports several experiments, but we shall consider only the data of Experiment I. Six groups were run. Two groups employed responses A_1 and A_2 and reinforcing events E_1 and E_2. The groups were denoted (70–30) and (60–40); the first number indicates the value of π, and the second the value of $1-\pi$. Asymptotic predictions for these groups are given by (4). The other groups involved three responses and were denoted (70–15–15), (70–20–10), (60–20–20), and (60–30–10); the first number indicates the value of π_1, the second the value of π_2, and the third the value of π_3. Asymptotic predictions for these groups are given by (8) for s equal to 1 and by (9) for s equal to 2.

The predicted values for s equal to 1 and 2 are presented in Table 1

TABLE 1

Predicted and Observed Asymptotic Proportions for the Gardner Data

Group	$P(A_1)$			$P(A_2)$			$P(A_3)$		
	Obs.	Predicted		Obs.	Predicted		Obs.	Predicted	
		s=1	s=2		s=1	s=2		s=1	s=2
60-40	.618	.600	.631	.382	.400	.369	----	----	----
60-30-10	.684	.600	.658	.235	.300	.267	.081	.100	.075
60-20-20	.676	.600	.667	.162	.200	.166	.162	.200	.166
70-30	.721	.700	.753	.279	.300	.279	----	----	----
70-20-10	.798	.700	.773	.129	.200	.156	.073	.100	.071
70-15-15	.802	.700	.800	.099	.150	.100	.099	.150	.100

along with Gardner's observed proportions over trials 286–450. Overall, the predictions for $s = 2$ give a fairly good account of the data. However, for comparable experimental procedures and equipment, one would hope that the number of response alternatives would not affect the estimated value of s. Unfortunately this invariance in s is not perfectly reflected in these data. For example, the predicted value of $P(A_1)$ for $s = 2$ is slightly high for the two-response groups and somewhat low for the three-response groups. Of course, this could be a statistical artifact, and a satisfactory answer would depend on a more detailed analysis of the sequential data.

There are several general comments to be made concerning these predictions. First of all, for s greater than one the predicted value of $P(A_1)$ in the (70–30) group is less than the predicted value of $P(A_1)$ for groups (70–15–15) and (70–20–10); similarly, the predicted value of $P(A_1)$ for the (60–40) group is less than $P(A_1)$ for groups (60–20–20) and (60–30–10). This result holds in general for the noncontingent reinforcement model: if the A_1 response is reinforced with some specified probability greater than $\frac{1}{2}$, then (for a fixed s greater than one) the prediction for $P(A_1)$ increases as a function of the number of alternative responses. Further, $P(A_1)$ approaches unity as s becomes large, independent of the number of alternative responses. Another result can be established for the three-response noncontingent model. Let $\pi_1 > \frac{1}{2}$, $\pi_2 \geq \pi_3$, and define $\delta = \pi_2 - \pi_3$. Then we can prove for fixed values of π_1 and s (where $s > 1$) that $P(A_1)$ increases as δ approaches 0.

REFERENCES

[1] Atkinson, R. C. A generalization of stimulus sampling theory. Tech. Rep. No. 29, Contr. Nonr 225(17), Appl. Math. and Statist. Laboratory, Stanford Univ., 1960.

[2] Estes, W. K. Toward a statistical theory of learning. *Psychol. Rev.*, 1950, **57**, 94-107.

[3] Estes, W. K. and Burke, C. J. A theory of stimulus variability in learning. *Psychol. Rev.*, 1953, **60**, 276-286.
[4] Gardner, R. A. Probability-learning with two and three choices. *Amer. J. Psychol.*, 1957, **70**, 174-185.
[5] Suppes, P. and Atkinson, R. C. *Markov learning models for multiperson interactions.* Stanford: Stanford Univ. Press, 1960.

Manuscript received 7/18/60

Revised manuscript received 11/28/60

PSYCHOMETRIKA—VOL. 26, NO. 3
SEPTEMBER, 1961

STATISTICAL METHODS FOR A THEORY OF CUE LEARNING

FRANK RESTLE*
MICHIGAN STATE UNIVERSITY

A theory of cue learning, which gives rise to a system of recurrent events in Feller's sense, is analyzed mathematically. The distribution of total errors and sampling distribution of mean errors are derived, and the learning curve is investigated. Maximum likelihood estimates of parameters and sampling variances of those estimates are derived. Likelihood ratio tests of the usual null hypotheses and approximate tests of goodness of fit of substantive hypotheses are developed. The distinguishing characteristic of these tests is that they are concerned with meaningful parameters of the learning process.

Consider a set H of strategies (consistent patterns of responses in a cue-learning experiment) of which a subset C always leads to a correct response, a subset W always leads to a wrong response, and the remainder, I, leads to correct and wrong responses at random. Let c, w, and i represent the proportion of strategies of each type, $c + w + i = 1$. Imagine that the subject selects one strategy at random from H and responds accordingly. If the response is correct, suppose that the subject continues with the same strategy; if the response is in error, the subject chooses at random from H (sampling with replacement) and obtains a new strategy for the next trial. The next response is based on this new strategy.

In such a model, errors are uncertain recurrent events [18] in Feller's sense. Quoting Feller, "The essential property of recurrent events is that after every occurrence of $\mathcal{E}$ the trials start from scratch. This means that all events preceding an occurrence of $\mathcal{E}$ should be statistically independent of all subsequent events" ([12], pp. 239–240). Furthermore, "... if $\mathcal{E}$ occurs at trial m, then the conditional probability that it occurs again at the trial number $m + n$ equals the probability that $\mathcal{E}$ occurs at the nth trial" ([12], p. 240).

Let f_j be the probability that $\mathcal{E}$ occurs for the first time at the jth trial. Then, for the learning model in question, it has been shown that

$$f_1 = w + \tfrac{1}{2}i, \qquad (1)$$

$$f_j = (\tfrac{1}{2})^j\, i \qquad (\text{for } j > 1).$$

These results were generalized in [18] by showing that the same process also arises if the subject uses strictly random samples of strategies with a certain

*This research was facilitated by the writer's tenure as Faculty Research Fellow, Social Science Research Council, 1959–1961.

procedure for eliminating wrong strategies, but without restriction on the size of the sample.

The purpose of the present paper is to establish statistical properties of this model so as to permit prediction of the details of data, and to lead to efficient estimates of parameters and statistical tests of a broad class of hypotheses. Some of the main statistical procedures are illustrated by examples.

Statistical Properties of the Data

For present purposes, the data for each subject consist of a sequence of correct and wrong responses. The responses are summarized by an indicator variable. Let $\mathbf{X}_{\alpha,n}$ be a random variable which takes the value 1 if subject α makes an error on trial n, and the value 0 otherwise. We assume that all subjects have the same parameters c, w, and i.

THEOREM 1. *Distribution of Total Errors.* Let

$$\mathbf{T}_\alpha = \sum_{n=1}^{\infty} (\mathbf{X}_{\alpha,n})$$

be the total errors made by subject α. Then

$$\Pr\{\mathbf{T}_\alpha = k\} = (1 - c)^k c. \tag{2}$$

PROOF. After any error the probability of at least one more error is

$$\sum_{j=1}^{\infty} (f_j) = w + i = 1 - c.$$

Hence the probability of no more errors is c. The probability that $\mathbf{T}_\alpha = k$ is Pr (at least one error) · Pr (at least one more error) · . . . (k times) . . . · Pr (no more errors) $= (1 - c)^k c$.

Theorem 1 gives a geometric distribution of error scores. This distribution has mean

$$E(\mathbf{T}_\alpha) = (1 - c)/c, \tag{3}$$

and variance

$$\text{Var}(\mathbf{T}_\alpha) = (1 - c)/c^2. \tag{4}$$

THEOREM 2. *Sampling Distribution of Mean Errors.* Let

$$\mathbf{T}_N = \sum_{\alpha=1}^{N} (\mathbf{T}_\alpha)$$

be the total errors made by a random sample of N subjects. Then

$$\Pr(\mathbf{T}_N = X) = \binom{X + N - 1}{N - 1}(1 - c)^X c^N. \tag{5}$$

PROOF. This is a well-known result, i.e., the sum of N random variables each of which has the same geometric distribution, itself has the negative binomial (Pascal) distribution given above ([12], p. 218).

This means that the mean errors per subject of samples of N subjects, $\bar{\mathbf{T}}_N$, will have the distribution

$$\Pr(\bar{\mathbf{T}}_N = X/N) = \binom{X + N - 1}{N - 1}(1 - c)^X c^N, \tag{6}$$

which is an explicit formula for the sampling distribution of the mean errors. This distribution has mean

$$E(\bar{\mathbf{T}}_N) = (1 - c)/c \tag{7}$$

and variance

$$\text{Var}(\bar{\mathbf{T}}_N) = (1 - c)/Nc^2. \tag{8}$$

THEOREM 3. *Markovian Property of Error Scores.* If one selects all subjects who make at least A (> 0) errors and asks the distribution of additional errors, this distribution is identical with the original distribution of errors. More exactly,

$$\Pr(\mathbf{T}_\alpha - A = y \mid \mathbf{T}_\alpha \geq A) = \Pr(\mathbf{T}_\alpha = y). \tag{9}$$

PROOF. This follows at once from the fact that errors are recurrent events; the system has exactly the same probabilities after the Ath error as it had originally, and hence the same distribution of errors.

This theorem was suggested by Bower's related theorem [4]. It can be tested directly by inspection of the conditional distributions and is a strong test of the correctness of the model. It is, however, equivalent to the geometric distribution of errors.

Generalization of Experimental Conditions

The above theorems are proved on the assumption that irrelevant strategies are rewarded independently and at random with probability one-half on each trial. It is sometimes difficult to ensure that all irrelevant strategies are so rewarded. Suppose that there are N_i irrelevant strategies, thought of as being equally potent. At any trial n, some number $N_{n,0}$ of the irrelevant strategies will lead to an error. Another disjoint set containing $N_{n,1}$ strategies will lead to a correct response on trial n and an error on trial $n + 1$. In general there exists a set (possibly empty) of irrelevant strategies, numbering $N_{n,j}$, which lead to correct responses on trials n, $n + 1$, $\cdots$, $n + j - 1$, and lead to an error on trial $n + j$.

THEOREM 4. If for all n,

$$\sum_{j=1}^{\infty} (N_{n,j}) = N_i ,$$

then the conclusions of Theorems 1–3 follow.

PROOF. Let $f_{n,j}$ be the conditional probability of an error on trial $n + j$ and correct responses on all trials between n and $n + j$, given an error on trial n. Now, $f_{n,1} = w + i(N_{n,1}/N_i)$, and for all j greater than 1, $f_{n,j} = i(N_{n,j}/N_i)$. The probability that the subject will ever make another error, given an error on trial n, is the sum of $f_{n,j}$ which, by the hypothesis of the theorem, is $w + i = 1 - c$ for all n. From this point on, proof of the theorem is straightforward.

This result is particularly interesting for experiments in which some strategies (other than the correct ones) are made correct on a proportion other than half of the trials, a procedure called "ambiguity of cues" by some experimenters [13]. Theorem 4 states that the total error score is unaffected by ambiguity of irrelevant cues, provided only that the irrelevant strategies eventually lead to error. In testing the theorem experimentally, it is essential that the subjects be trained to an extremely rigorous criterion to ensure that (almost) all the errors to be made are squeezed out.

Effects of Perseveration

A variation of the theory, mentioned in other publications [17, 18], makes the following assumption: when a correct response is made the subject continues with the same strategy, and when an error is made the subject stays with the same strategy with probability $1 - r$, and resamples (with replacement) with probability r.

This variation is not quite a system of recurrent events. The probability of zero errors is c, as in the simple model above, but the probability of making no more errors after any error is only rc, the joint probability that the subject resamples (r) times the conditional probability that he selects a correct strategy given that he resamples (c). The following results are derived by methods analogous to those used to prove Theorems 1–4, and are written down without proof.

Distribution of Total Errors

$$\Pr(\mathbf{T}_a = k) = \begin{cases} c & (\text{for } k = 0) \\ (1 - c)(1 - rc)^{k-1}rc & (\text{for } k > 0). \end{cases} \tag{10}$$

This distribution has mean and variance

$$E(\mathbf{T}_a) = (1 - c)/rc, \tag{11}$$

$$\text{Var}(\mathbf{T}_a) = (1 - c)(1 - rc + c)/(rc)^2. \tag{12}$$

If we delete all subjects who make no errors, we obtain the following conditional statistics:

$$\Pr(\mathbf{T}_\alpha = k \mid \mathbf{T}_\alpha \geq 1) = (1 - rc)^{k-1}rc, \tag{13}$$

which has mean and variance

$$E(k \mid \mathbf{T}_\alpha \geq 1) = 1 + \frac{1 - rc}{rc} = \frac{1}{rc}, \tag{14}$$

$$\operatorname{Var}(k \mid \mathbf{T}_\alpha \geq 1) = (1 - rc)/(rc)^2. \tag{15}$$

It at once follows that the conditional sampling distribution (again deleting all subjects who make no errors) is given by

$$\Pr(\bar{\mathbf{T}}_N = X/N \mid \text{all } \mathbf{T}_\alpha \geq 1) = \binom{X + N - 1}{N - 1}(1 - rc)^X(rc)^N, \tag{16}$$

which has mean and variance of

$$E(\bar{\mathbf{T}}_N \mid \text{all } \mathbf{T}_\alpha \geq 1) = (1 - rc)/rc, \tag{17}$$

$$\operatorname{Var}(\bar{\mathbf{T}}_N \mid \text{all } \mathbf{T}_\alpha \geq 1) = (1 - rc)/(rc)^2N.$$

Effects of Conjunctive Compound Correct Strategies

Some problems require the subject to acquire more than one correct strategy. Provided that several strategies are sampled and fixated independently, each with the same probability of selection c, the total-error distribution for individual subjects will be the distribution of the sum of the errors accumulated for each required correct strategy. This distribution is given by Theorem 2, in which $\mathbf{T}_N$ is to be interpreted as the number of errors made by a given subject in a problem which requires N correct strategies all to be selected.

The Learning Curve

We now return to the simple model in which irrelevant strategies have probability one-half of leading to a correct response, there is no perseveration, and a single correct strategy suffices to solve the problem.

An interesting characteristic of the model is the complexity of the expression for the learning curve. Let $u_n = \Pr(\mathbf{X}_{\alpha,n} = 1)$ be the probability of an error on trial n. Then one has the fundamental equation of recurrent events ([12], ch. 12) as a recursive expression for u_n.

$$u_n = f_n + u_1f_{n-1} + u_2f_{n-2} + \cdots + u_{n-2}f_2 + u_{n-1}f_1. \tag{18}$$

Substituting the values of f_i from (1) this becomes

$$u_n = i(\tfrac{1}{2})^n + u_1i(\tfrac{1}{2})^{n-1} + \cdots + u_{n-1}(w + \tfrac{1}{2}i), \tag{19}$$

$$u_{n+1} = i(\tfrac{1}{2})^{n+1} + u_1 i(\tfrac{1}{2})^n + \cdots + u_n(w + \tfrac{1}{2}i). \tag{20}$$

Subtracting (19) from (20) and simplifying, we obtain the second-order difference equation

$$u_{n+1} = (\tfrac{1}{2} + \tfrac{1}{2}i + w)u_n - \tfrac{1}{2}wu_{n-1} . \tag{21}$$

Equation (21) is resolved by considering paired values of (u_{n+1} , u_n) for $n = 1, 2, \cdots$. A matrix equation arises,

$$(u_{n+1} , u_n) = (u_n , u_{n-1}) \begin{bmatrix} \frac{1}{2} + \frac{1}{2}i + w & 1 \\ -\frac{1}{2}w & 0 \end{bmatrix},$$

and since one can obtain u_1 and u_2 directly,

$$(u_{n+1} , u_n) = (u_2 , u_1) \begin{bmatrix} \frac{1}{2} + \frac{1}{2}i + w & 1 \\ -\frac{1}{2}w & 0 \end{bmatrix}^{(n-1)}. \tag{22}$$

Unfortunately the solution to (22) is so complex in form as to be valueless for the usual purposes. However, the actual learning curve can be investigated indirectly in several ways, which give a good idea of its behavior. For a given value of c, the expected total number of errors is $(1 - c)/c$ from Theorem 1. It is obvious, therefore, that

$$\sum_{n=1}^{\infty} (u_n) = (1 - c)/c. \tag{23}$$

Hence, we know the sum of the u_n . Furthermore, for fixed c, one has an array of possible learning curves depending on how $1 - c$ is distributed over w and i. The two extreme cases are $w = 0$ and $w = 1 - c$. For $w = 0$ the recursive equation (21) becomes a first-order difference equation,

$$\begin{aligned} u_{n+1} &= (\tfrac{1}{2} + \tfrac{1}{2}i)u_n \\ &= (1 - \tfrac{1}{2}c)u_n , \end{aligned} \tag{24}$$

which has the solution

$$u_{n+1} = u_1(1 - \tfrac{1}{2}c)^n. \tag{25}$$

When $w = 0$, $u_1 = f_1 = \frac{1}{2} - \frac{1}{2}c$, so that

$$u_{n+1} = (\tfrac{1}{2} - \tfrac{1}{2}c)(1 - \tfrac{1}{2}c)^n. \tag{26}$$

At the other extreme, for $w = 1 - c$ and $i = 0$, the situation simplifies because $f_j = 0$ for all $j > 1$, whereas $f_1 = w = 1 - c$. Equation (21) becomes

$$u_n = u_{n-1}(1 - c).$$

In this case $u_1 = 1 - c$ so that

$$u_n = (1 - c)^n. \tag{27}$$

Equations (26) and (27) differ in u_1, the initial probability of an error, and in the rate parameter which is $c/2$ with irrelevant but no wrong strategies, and c with wrong but no irrelevant strategies. With only wrong strategies, the initial probability of an error is high but is reduced rapidly; with only irrelevant strategies the initial probability of an error is lower but is reduced less rapidly. When both w and i are different from zero the learning curve is not exponential as it is in the extreme cases, and one cannot be sure that it lies between the two extreme curves.

Estimates of Parameters

An important statistical step in any learning model is the estimation of parameters. Fortunately there exist simple maximum likelihood estimators of the parameters c and w (and hence of i, since $i = 1 - c - w$) in the present model.

A separate estimate of c is easily derived from (3). Consider that some group of N subjects make a total of X, or a mean of X/N, errors. The likelihood of such an outcome, as a function of the possible values of the parameter c, is

$$L \text{ (set of data with } X \text{ errors; } c) = (1 - c)^X c^N, \tag{28}$$

which is the probability of any particular distribution of X errors over N people. The value of L is to be maximized with respect to c. As usual it is more convenient to maximize log (L). Taking the derivative with respect to c and setting it equal to zero,

$$0 = \frac{d}{dc} \log (L) = \frac{d}{dc} [X \log (1 - \hat{c}) + N \log (\hat{c})] = -\frac{X}{1 - \hat{c}} + \frac{N}{\hat{c}},$$

whence

$$\hat{c} = \frac{N}{N + X} = \frac{1}{1 + \bar{T}}. \tag{29}$$

This is a natural estimate. If one used the method of moments using the fact that $E(\mathbf{T}) = (1 - c)/c$ and equating $E(\mathbf{T})$ with $\bar{T}$, one would obtain the same result.

With large samples, the sampling variance of a maximum likelihood estimate (provided the estimates are normally distributed) can be calculated by the formula

$$\text{Var } (\hat{c}) = \frac{-1}{E\left(\frac{d^2}{dc^2} \log L\right)}, \tag{30}$$

which, in the present case, is

$$\text{Var } (\hat{c}) = \frac{-1}{E\left(-\frac{X}{(1-c)^2} - \frac{N}{c^2}\right)}.$$

Since $E(X) = N(1 - c)/c$,

$$\text{Var } (\hat{c}) = c^2(1 - c)/N. \tag{31}$$

This may be approximated by substituting $\hat{c}$ for c, and, if in addition we substitute $X/(X + N)$ for $\hat{c}$, we obtain the estimate

$$\text{Approximate Var } (\hat{c}) = \frac{NX}{(X + N)^3}. \tag{32}$$

Further properties of the estimate are discussed below.

For some purposes it will be desirable to obtain estimates of both c and w. Consider the sequence $\{\mathbf{X}_{\alpha,n}\}$ for each subject α. Provided that $c > 0$, this sequence will have an infinitely long terminal string of 0's, and a last error (which may be at "trial zero," i.e., no error at all). We divide each protocol into a presolution phase (up to the last error) and a terminal phase (the terminal string of 0's).

The conditional probability of an error following an error, $P(1 \mid 1)$, in the presolution phase is $(1 - c + w)/2$, and the conditional probability of a correct response following an error, $P(0 \mid 1) = (1 - c - w)/2$, except for the correct response following the last error which is, however, part of the terminal phase. The other conditional probabilities, $P(1 \mid 0)$ and $P(0 \mid 0)$, are both equal to one-half in the presolution phase. In the presolution phase $P(1 \mid 0)$ and $P(0 \mid 0)$ are not functions of the parameters and can be disregarded in estimation.

Consider all errors made by subject α, which are T_α in number, and add "trial zero" to obtain $T_\alpha + 1$ opportunities to learn. Suppose that $M_1^{(\alpha)}$ of these occasions are followed by errors and $M_0^{(\alpha)}$ are followed by correct responses. Then, disregarding $P(1 \mid 0)$, $P(0 \mid 0)$, and the binomial coefficients which do not depend upon the parameters, we find

$$\begin{aligned} L_\alpha &= (P(1 \mid 1))^{M_1^{(\alpha)}} \cdot (P(0 \mid 1))^{M_0^{(\alpha)}-1} c \\ &= (\tfrac{1}{2}(1 - c + w))^{M_1^{(\alpha)}} (\tfrac{1}{2}(1 - c - w)^{M_0^{(\alpha)}-1} c. \end{aligned}$$

The likelihood of a whole set of data for N subjects is

$$L_N = \prod_{\alpha=1}^{N} (L_\alpha).$$

If we let M_1 be the total number of errors following errors, and M_0 be the total number of correct responses following errors in the whole group, it follows that

$$L_N = \left(\frac{1 - c + w}{2}\right)^{M_1}\left(\frac{1 - c - w}{2}\right)^{M_0 - N} c^N. \tag{33}$$

Maximum likelihood estimators of c and w are obtained by differentiating (33) with respect to c and w, setting the two derivatives equal to zero, and solving the resulting pair of equations simultaneously. The logarithm of L_N is used.

$$\log (L_N) = M_1 \log (\tfrac{1}{2}) + M_1 \log (1 - c + w) + (M_0 - N) \log (\tfrac{1}{2}) + (M_0 - N) \log (1 - c - w) + N \log (c). \tag{34}$$

$$\frac{\partial}{\partial c} \log (L_N) = \frac{-M_1}{1 - \hat{c} + \hat{w}} + \frac{-(M_0 - N)}{1 - \hat{c} - \hat{w}} + \frac{N}{\hat{c}} = 0. \tag{35}$$

$$\frac{\partial}{\partial w} \log (L_N) = \frac{M_1}{1 - \hat{c} + \hat{w}} + \frac{-(M_0 - N)}{1 - \hat{c} - \hat{w}} = 0. \tag{36}$$

From (36),

$$\hat{w} = \frac{M_1 - (M_0 - N)}{M_1 + (M_0 - N)} (1 - \hat{c}).$$

If this is substituted into (35) and the result simplified,

$$\hat{c} = N/(M_1 + M_0). \tag{37}$$

Recalling that $M_1 + M_0$ is the total errors made by all subjects plus one "trial zero" for each subject, it is seen that (37) agrees with (29). Substituting (37) back into (36),

$$\hat{w} = \frac{M_1 - M_0 - N}{M_1 + M_0}. \tag{38}$$

If we let $\bar{M}_1$ and $\bar{M}_0$ be the means of $M_1^{(\alpha)}$ and $M_0^{(\alpha)}$,

$$\hat{w} = \frac{\bar{M}_1 - \bar{M}_0 - 1}{\bar{T} + 1}. \tag{39}$$

Equations (37) and (38) constitute maximum likelihood estimators for the main parameters of the model.

The estimator of c is biased, but an unbiased estimator can be constructed. Consider samples of size N. From (6),

$$\Pr (\mathbf{T}_N = X/N) = \binom{X + N - 1}{N - 1}(1 - c)^X c^N.$$

If one observes X errors then the estimate of c is $N/(X + N)$. Hence the expected value of $\hat{c}$ is

$$E(\hat{c}) = \sum_{X=0}^{\infty} \frac{N}{X + N} P(\bar{\mathbf{T}}_N = X/N)$$
$$= \sum_{X=0}^{\infty} \binom{X + N - 1}{N - 1} (1 - c)^X c^N \left(\frac{N}{X + N}\right).$$

An algebraic rearrangement of binomial coefficients and c^N gives

$$E(\hat{c}) = c \sum_{X=0}^{\infty} \frac{N}{X + N} \cdot \frac{X + N - 1}{N - 1} \cdot \binom{X + N - 2}{N - 2} (1 - c)^X c^{N-1}.$$

The binomial coefficients and terms to the right are the terms of a Pascal distribution, hence the sum is the mean of $N(X + N - 1)/(X + N)(N - 1)$. Since the ratio $(N - 1)/N$ will be smaller than $(X + N - 1)/(X + N)$ for all $X > 0$, the sum is the mean of a set of numbers all but one of which are greater than 1, and the estimate is biased toward the larger side. An unbiased estimator is

$$\hat{c} = \frac{N - 1}{X + N - 1}. \tag{40}$$

Maximum likelihood estimators are also available when one invokes perseveration. Let N_0 be the number of subjects who make zero errors. Then the likelihood function is

$$L = c^{N_0}(1 - c)^{N-N_0}(1 - rc)^{X-N+N_0}(rc)^{N-N_0}. \tag{41}$$

Differentiating the logarithm of (41) with respect to c and r and setting the two partial derivatives equal to zero one obtains

$$\frac{N_0}{\hat{c}} - \frac{N - N_0}{1 - \hat{c}} - \frac{r(X - N + N_0)}{1 - \hat{r}\hat{c}} + \frac{r(N - N_0)}{\hat{r}\hat{c}} = 0. \tag{42}$$

$$\frac{X - N + N_0}{1 - \hat{r}\hat{c}} = \frac{N - N_0}{\hat{r}\hat{c}}. \tag{43}$$

From (43) we see that the last two terms of (42) cancel, whence from (42),

$$\hat{c} = N_0/N, \tag{44}$$

and

$$\hat{r} = N(N - N_0)/N_0X. \tag{45}$$

The following equations give sampling variances of the above estimates:

$$\text{Var}(\hat{c}) = \frac{1}{\left[\frac{N - N_0}{(1 - c)^2} + \frac{N}{c^2} + \frac{r(1 - c - Nrc + N_0rc)}{c(1 - rc)^2}\right]}; \tag{46}$$

$$\text{Var}(\hat{r}) = \frac{1}{\left[\frac{rc(X - N + N_0)}{(1 - rc)^2} + \frac{N - N_0}{r^2}\right]}. \tag{47}$$

Both of these approximate values of the variance can be estimated from the data. Their validity depends on the hypothesis that the sampling distribution of the estimate is normal, and they are usable for large N.

Statistical Tests of Hypotheses

In the present model the sampling distribution of mean errors (5) is asymptotically normal, so that one can use the usual statistical procedures to compare two means. However, it would be better to perform statistical tests directly in terms of the parameters of the model and estimates thereof. In particular, one may wish to test a certain quantitative law, which is almost certain to be stated in its simplest and most basic form in terms of parameters such as c, w, i, or r rather than descriptive statistics of the data. In the following discussion some such statistical tests are derived, and some problems are discovered.

For some simple situations, likelihood ratio tests can be constructed. From the discussions above it is known that the likelihood of any particular set of data in which N subjects make a total of X errors is $(1 - c)^X c^N$, which can be considered as a function of the possible values of c. In a likelihood ratio test one maximizes the likelihood over a restricted parameter subspace (the null hypothesis) Ω_0 , and also over the entire space of logical possibilities, Ω. The ratio of these two likelihoods is called λ and is given by

$$\lambda = \frac{\max_{\theta \epsilon \Omega_0} L(X, \theta)}{\max_{\theta \epsilon \Omega} L(X, \theta)}, \tag{48}$$

where θ is some parameter such as c, w, r, etc. With large samples and under quite general conditions the value of $-2 \ln \lambda$ is distributed approximately as χ^2, provided the null hypothesis is true. The degrees of freedom of the χ^2 are calculated by seeing how many more parameters are estimated in Ω than in Ω_0 .

Test of an Hypothetical Parameter c_0

Consider a test of the hypothesis that $c = c_0$. Then Ω_0 is merely the set whose only member is c_0 , whereas Ω is the whole interval from 0 to 1. The ratio of the likelihoods is

$$\lambda = \frac{(1 - c_0)^X c_0^N}{\left(\frac{X}{X + N}\right)^X \cdot \left(\frac{N}{X + N}\right)^N}, \tag{49}$$

from which one can calculate $-2 \ln \lambda$ which, in the event that c_0 actually is the parameter value, will have approximately a χ^2 distribution with one degree of freedom. This test corresponds to the simple t test of a fixed parameter in ordinary statistical methods, except that the present test uses the assumptions of the model and tests a meaningful parameter.

Test that Two Sets of Data Have the Same Parameter

In a simple two-sample test the universe Ω is the unit square and the null hypothesis subspace Ω_0 is the diagonal line. Computation of the two likelihoods is straightforward; the maximum likelihood estimates of c_1 and c_2, maximizing over the whole space, are $N_1/(N_1 + X_1)$ and $N_2/(N_2 + X_2)$, from which one can calculate the denominator of the likelihood ratio. The maximum likelihood estimate of c on the diagonal line is

$$(N_1 + N_2)/(N_1 + N_2 + X_1 + X_2)$$

from which the numerator of the likelihood ratio is computed. Provided the sample sizes are sufficiently large, $-2 \ln \lambda$ will have approximately a χ^2 distribution with one degree of freedom.

More Complex Hypotheses about c

The main use of a mathematical learning theory like the present one, in experimental applications, is to mediate predictions about the solution of one problem based on information gathered using another, related problem. Such predictions are almost never perfect, and an imperfect prediction raises the question as to whether the discrepancy can be attributed to sampling deviations.

An example of a prediction from such a model is "additivity of cues" [16, 18, 19]. One has a problem in which a set C_1 of strategies are correct and the set C_2 are irrelevant. A second problem makes the set C_2 correct and C_1 irrelevant. A third problem has the set $C_1 \cup C_2$ correct; all other strategies are presumed to be the same in all three problems. In this experiment, if C_1 and C_2 are disjoint sets, the proportions of correct strategies in the three groups are related by the equation

$$c_1 + c_2 = c_3 \,. \tag{50}$$

The statistical problem is to decide whether a given set of data are in agreement with (50).

One approach to such a problem is to find the maximum likelihood estimates of c_1 and c_2 which satisfy (50), and another set of maximum likelihood estimates of c_1, c_2, and c_3. If this can be done a likelihood ratio test can be constructed at once. In the case of the additivity-of-cues experiment the likelihood ratio test can be performed only by solving a pair of fourth-degree equations. In other experiments with more complicated hypotheses,

it may be virtually impossible to determine the maximum likelihood estimates.

In the case of additivity of cues, a feasible method is to estimate c_1 and c_2 from performance on the first two problems. One has, of course, maximum likelihood estimates, and from (31) one has an estimate of the variance of the parameter estimates. If the sample size is sufficient, we can assume that the sampling distributions of the estimates are approximately normal.

Now if we apply (50) we add the estimates of c_1 and c_2 ; this prediction will also be normally distributed (so far as sampling deviations are concerned) with variance equal to the sum of the variances of c_1 and c_2 . Group 3 also yields an estimate of c_3 and an estimate of the sampling variance of the parameter estimate, which we may assume to have an approximately normal distribution. The statistical test is then the comparison of two normal distributions which, under the null hypothesis, have equal means. The test is dependable only with large samples, but one is not likely to attempt serious evaluation of such a quantitative hypothesis except by a sizable experiment.

The method given above overcomes the difficulty faced in earlier tests of the additivity hypothesis [16], in which one merely estimated c_1 and c_2 , added the estimates, and then from that sum computed a prediction of, say, mean errors to be made by group 3. The test consisted of asking whether the obtained mean errors of group 3 differed significantly from this prediction when one considers only the estimated sampling variance of the mean of group 3. This test is overly stringent since it does not take account of the sampling variance of the prediction, which depends upon sampling variance in groups 1 and 2. The present method, while not entirely exact, seems to correct the excessive stringency of the earlier method. For other approaches to the problem, see [19].

Other empirical hypotheses which one may wish to test require more complicated statistical procedures. One may, for example, consider the additivity of irrelevant hypotheses [3]; in problems with various numbers (I) of equally potent irrelevant dimensions, one can conclude (with suitable restrictions) that

$$c_I = \frac{c_0}{1 + aI} ,$$

where a is a parameter representing the potency of an irrelevant dimension. In order to test this hypothesis one must first estimate the parameter a. Investigations of the likelihood expressions indicate that maximum likelihood estimates of such parameters will be quite difficult to obtain.

One simple approach, which can always be applied, is to compute estimates of c for each of the groups, and also estimate the sampling variance of these estimates. The hypothetical equation can then be investigated by curve-fitting procedures. If one assumes that the sampling distribution of $\hat{c}$

is normal, least squares, weighted least squares, or even graphical estimation methods can be used, taking account of the sampling variance of the parameter estimates being fitted. When a satisfactory fit of the hypothetical equation is obtained, one has a "predicted" parameter $\dot{c}$ for each group. Using the fact that

$$\text{Var}\,(\hat{c}) = c^2(1 - c)/N,$$

one can calculate a theoretical variance about each predicted point. A reasonable test of whether the fitted curve is satisfactory can then be obtained by taking

$$D^2 = \sum_{\substack{\text{all} \\ \text{groups}}} \frac{(\hat{c}_i - \dot{c}_i)^2}{\sqrt{\text{Var}\,(\hat{c}_i)}}.$$

Provided that the sampling distribution of parameter estimates is normal, and that $\dot{c}_i$ is for each group the correct parameter value, the statistic D^2 should have a χ^2 distribution. The number of degrees of freedom would be the total number of groups minus the number of independent parameters estimated to obtain the values of $\dot{c}_i$. The most serious problem in applying this approach is the difficulty of making good estimates of such theoretical parameters as a. That is, one requires a good solution to the curve-fitting problem before the results of the statistical test are clear. Notice that D^2 distributes as χ^2 only if the $\dot{c}$ are the correct parameter values. With an unfortunate job of curve fitting one might reject too often a theoretical equation which is correct.

Discussion

A stochastic model of learning can be used to investigate substantive questions which go beyond the particular assumptions of the model. The first step in using such a model is to assess whether the model fits the data adequately. Besides the ordinary goodness of fit statistics, it is desirable to study a variety of different statistics of the data [7] to find the weak points in correspondence between model and data. However, one should not stop with goodness of fit. If one has a good model, it will often be easier to state quantitative hypotheses in terms of the parameters of the model than directly in terms of statistics of the raw data. Hypotheses of this kind have already been offered in the literature [3, 8, 9, 14], and the present results are helpful in testing such hypotheses. There are other quantitative laws which are currently stated in terms of statistics of the data [5, 6, 10, 15] and can be tested without the special statistical results developed in this paper. However, the theoretical significance of experimental predictions can often be increased by making more and more precise and detailed predictions—and the kind of statistical methods developed above are most useful when one goes into great detail.

It should be borne in mind that the statistical methods developed in this paper arise from the assumptions of a particular model of learning and cannot properly be applied unless that model fits the situation. The restriction is severe in the use of the chi-square tests of hypotheses. for these tests assume (i) that there are no individual differences in the parameter c and (ii) that all learning processes are independent. If each subject learns more than one thing (for example, a whole list of nonsense syllables or a sequence of two-choice problems) it is quite likely that observations will correlate differently for between- and within-subjects data, and the model will not fit the agglomerated data. The result can invalidate the statistical tests given above. It is emphasized tLat the statistical methods and tests proposed in this paper are not "general purpose" procedures, but are designed to meet special needs when it has already been established that the general model agrees closely with the stochastic structure of the data.

However, the strategy-selection model used, since it yields a simple Markovian structure, is closely related to a variety of other models of learning [1, 2, 4, 11]. It is hoped that the structure given above can be generalized to fit a variety of related learning models.

The theorems of this paper were derived mainly through a single conceptual trick—that of considering the number of errors made without attention to exactly when they are made. It was found that no useful expression for the general learning curve could be derived. However, by asking the probability that another error would be made at all, irrespective of the specific trial on which it would be made, it was possible to derive the distribution of total error scores, the sampling distribution of mean error scores, maximum likelihood estimators of the main parameters, and some statistical tests. The mathematical method is a simple application of Feller's theory of recurrent events [12], and may be useful in mathematical investigations of learning models.

REFERENCES

[1] Atkinson, R. C. A theory of stimulus discrimination learning. Tech. Rep. 1, Contr. Nonr 233(58), Dept. Psychol., Univ. California, Los Angeles, 1959.

[2] Atkinson, R. C. and Suppes, P. An analysis of a two-person interaction situation in terms of a Markov process. Tech. Rep. 9, Contr. Nonr 225 (17), Appl. Math. and Statist. Laboratory, Stanford Univ., 1957.

[3] Bourne, L. E., Jr. and Restle, F. Mathematical theory of concept identification. *Psychol. Rev.*, 1959, **66**, 278-296.

[4] Bower, G. H. Properties of the one-element model as applied to paired-associate learning. Tech. Rep. 31, Contr. Nonr 225 (17), Inst. for Math. Stud. in the Soc. Sci., Stanford Univ., 1960.

[5] Bush, R. R., Galanter, E., and Luce, R. D. Tests of the "Beta Model." In R. R.Bush and W. K. Estes (Eds.), *Studies in mathematical learning theory.* Stanford: Stanford Univ. Press, 1959. Pp. 381-399.

[6] Bush, R. R. and Mosteller, F. *Stochastic models for learning.* New York: Wiley, 1955.

[7] Bush, R. R. and Mosteller, F. A comparison of eight models. In R. R. Bush and W. K. Estes (Eds.), *Studies in mathematical learning theory*. Stanford: Stanford Univ. Press, 1959. Pp. 293-307.

[8] Estes, W. K. Statistical theory of spontaneous recovery and regression. *Psychol. Rev.*, 1955, **62**, 145-154.

[9] Estes, W. K. Statistical theory of distributional phenomena in learning. *Psychol. Rev.*, 1955, **62**, 369-377.

[10] Estes, W. K. Theory of learning with constant, variable, or contingent probabilities of reinforcement. *Psychometrika*, 1957, **22**, 113-132.

[11] Estes, W. K. Component and pattern models with Markovian interpretations. In R. R. Bush and W. K. Estes (Eds.), *Studies in mathematical learning theory*. Stanford: Stanford Univ. Press, 1959. Pp. 9-52.

[12] Feller, W. *Introduction to probability theory and its applications*. (1st ed.) New York: Wiley, 1950.

[13] Gormezano, I. and Grant, D. A. Progressive ambiguity in the attainment of concepts on the Wisconsin Card Sorting Test. *J. exp. Psychol.*, 1959, **55**, 621-627.

[14] Restle, F. Toward a quantitative description of learning set data. *Psychol. Rev.*, 1958, **65, 77-91.**

[15] Restle, F. A survey and classification of learning models. In R. R. Bush and W. K. Estes (Eds.), *Studies in mathematical learning theory*. Stanford: Stanford Univ. Press, 1959. Pp. 415-428.

[16] Restle, F. Additivity of cues and transfer in discrimination of consonant clusters. *J. exp. Psychol.*, 1959, **57**, 9-14.

[17] Restle, F. Note on the "hypothesis" theory of discrimination learning. *Psychol. Rep.*, 1960, **7**, 194.

[18] Restle, F. The selection of strategies in cue learning. *Psychol. Rev.* (in press).

[19] Trabasso, T. R. Additivity of cues in discrimination learning of letter patterns. *J. exp. Psychol.*, 1960, **60**, 83-88.

Manuscript received 11/11/60

Revised manuscript received 3/11/61

PSYCHOMETRIKA—VOL. 26, NO. 3
SEPTEMBER, 1961

THE USE OF EXTREME GROUPS TO TEST FOR THE PRESENCE OF A RELATIONSHIP

LEONARD S. FELDT
STATE UNIVERSITY OF IOWA

Experimenters in psychology frequently investigate the relationship between two variables by selecting extreme groups on the first measure and comparing their mean scores on the second. This paper considers the efficacy of this procedure from the criterion of the power of the statistical tests. Optimal cutting points for the extreme groups are defined, and the power of the difference approach is compared to that of significance tests for the product-moment correlation coefficient.

In the exploratory stages of the investigation of a psychological construct, a two-stage experimental design is frequently employed. In the first stage, a random sample of subjects from a hypothetical or real population is evaluated via a crude measure of the construct, and from the distribution of scores which results, an arbitrary definition is derived for "High" and "Low" subgroups. In the second stage, the high and low groups are exposed to one or several treatment conditions. It is hypothesized that if the initial classification was even moderately valid, the treatment should produce a different distribution of treatment criterion scores for the high and low subpopulations. Usually such a difference is assessed through a comparison of the means of the groups.

Examples of this design are quite common in the psychological literature. It was frequently used, for example, in the early studies of McClelland's Achievement Need [7]. It was also extensively employed in the preliminary validation of Taylor's Manifest Anxiety Scale [11]. In the latter studies the second stage often involved fairly complex and time consuming treatment conditions, such as eyelid conditioning. Thus, as is often the case, the use of extreme groups was prompted in part by the necessity to limit the total number of subjects.

An important consideration in such a design is the choice of upper and lower percentiles which define the extreme groups. Current experimental practice evidences considerable variation in this aspect of the design. Some investigators have employed extreme tenths or fifths, others have utilized upper and lower halves. In all cases the decision appears to have been quite arbitrary, or dictated by necessity. The primary purpose of this paper is to derive a definition of optimal extreme groups for investigations of this kind. A second purpose is to compare the efficiency of this design to the correlation approach which provides an obvious alternative procedure.

Definitions and Assumptions

In the following development, the initial classification variable, the validity of which is under test, will be designated as X. The criterion measure taken in the second stage of the experiment—strength of eyelid conditioning in the Taylor Scale example—will be designated Y. Measures X and Y will be assumed to give rise to a normal bivariate surface with correlation ρ_{xy} in the population of potential experimental subjects. The experiment itself consists of obtaining measure X on a random sample from the subject population, defining equal extreme subgroups on X, imposing treatments conditions and obtaining the Y criterion score on each subject, and finally testing the significance of the difference between Y means for the two groups via a t test.

In the definition of optimal extreme groups, the criterion employed will be the power of the final t test. For a given level of significance the power of this test is dependent upon three quantities: (i) the variability of the Y measures within each extreme group, (ii) the magnitude of the true difference between the Y means; and (iii) the size of the extreme groups. Each of these factors is functionally related to the percentiles chosen to define the groups. It is the nature of linear regression that the more extreme the subpopulations, the larger the difference between the Y means. A large difference is clearly advantageous, for it will result in a more powerful test than will a small difference. But it is also true that the more select the subpopulations, the smaller the number of subjects available. This, in turn, tends to increase the standard error of the difference. Thus, if only the upper and lower ten percent of subjects is employed, both the difference and the standard error of the difference will be larger than if upper and lower halves are used. The value of the variance of Y scores within the subpopulations will also vary with the "extremeness" of the subpopulations. The problem thus becomes one of deriving symmetrical upper and lower percentiles which will result in a combination of true difference, group size, and within-group variance that will yield the most powerful test.

Optimal Definition of Extreme Groups

For the t test of the difference between means of extreme samples let N represent the total number of subjects initially classified, p the proportion of subjects in each extreme group, $n = pN$ the number of subjects in each group, S^2 the sample variance, and the subscripts U and L the upper and lower groups respectively. It should be noted that in this situation N is fixed; the problem involves the determination of the optimal value of p. The t test for equal groups may be written as follows:

$$t = \frac{|\bar{Y}_U - \bar{Y}_L|}{\sqrt{\frac{S_U^2 + S_L^2}{n - 1}}}.$$

The power of this test is governed by the parameter ϕ, defined as follows [10]:

$$\phi = \frac{\mu_w}{\sqrt{2}\sigma_w}. \tag{1}$$

In this formula μ_w is the expected value or mean of a normally distributed variable, say w, and σ_w is its population standard deviation. In the present context $(\bar{Y}_U - \bar{Y}_L)$ is the normally distributed variable, $(\mu_U - \mu_L)$ is its expected value, and $\sigma_{\bar{Y}_U - \bar{Y}_L}$ is its population standard deviation. Thus

$$\phi = \frac{|\mu_U - \mu_L|}{\sqrt{2}\sigma_{\bar{Y}_U - \bar{Y}_L}} = \frac{|\mu_U - \mu_L|}{\sqrt{\frac{2(\sigma^2_{Y_U} + \sigma^2_{Y_L})}{pN}}}. \tag{2}$$

The value of σ_{Y_U} or σ_{Y_L} is given in [2] as $\sigma^2_Y(1 - c\rho^2)$. In this relationship, σ^2_Y is the variance of the total population on the treatment criterion, ρ is the population value of the linear correlation between the classification variable and the treatment criterion, and c is a constant dictated by the degree of "extremeness" of the upper and lower groups. The defining relationship for c is

$$c = 1 - \frac{\sigma^2_{x\ \text{extreme}}}{\sigma^2_{x\ \text{total}}}. \tag{3}$$

The second term on the right-hand side is the ratio of the variance within either subpopulation (they are equally variable) to the variance for the total population on the classification variable.

For the normal surface here assumed, the value of the ratio may be computed, by integration by parts, for any selected segment of the distribution [4]. Substitution of this result in (3) yields

$$c = \frac{z^2}{p^2} - \frac{xz}{p}, \tag{4}$$

where x is the positive or negative standard normal deviate defining the extreme groups, and z is the ordinate of the standard normal curve at x. It may also be noted that μ_{Y_U} and μ_{Y_L} are defined by the regression line of Y on X as follows:

$$\mu_{Y_U} = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(\mu_{X_U} - \mu_X);$$

$$\mu_{Y_L} = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(\mu_{X_L} - \mu_X).$$

The difference thus equals

$$\mu_{Y_U} - \mu_{Y_L} = \rho\frac{\sigma_Y}{\sigma_X}(\mu_{X_U} - \mu_{X_L}). \tag{5}$$

By appropriate integration, the values of μ_{X_U} and μ_{X_L} may be derived for any segment of a normal curve [4]. In the notation previously introduced,

the means for equal extreme segments are $\mu_X \pm (z/p)\sigma_X$. Upon substitution of these expressions into the previous equation, the difference becomes

$$\mu_{Y_U} - \mu_{Y_L} = \frac{2\rho\sigma_Y z}{p}. \tag{6}$$

Substitution of results (4) and (6) into (2) and the division of numerator and denominator by $2\sigma_Y$ yields the following expression for ϕ:

$$\phi = \frac{\rho\frac{z}{p}}{\sqrt{\frac{1 - \rho^2\left(\frac{z^2}{p^2} - \frac{xz}{p}\right)}{pN}}}. \tag{7}$$

To secure maximum power in the test of the difference between $\bar{Y}_U$ and $\bar{Y}_L$, ϕ must be maximized. This is accomplished by appropriate choice of x, and of z and p, which are both functionally related to x. Since the quantity N is fixed in the context of this problem, its value may be ignored. The condition for a maximum is obtained by setting the first derivative of ϕ with respect to x equal to zero. The process of taking this derivative and of solving the resultant equation involves some rather tedious algebra, and hence it will not be presented here. The end result, however, may be stated quite simply. The maximum value of ϕ occurs when

$$\rho^2\left(1 + x^2 - \frac{z^2}{p^2}\right) + \frac{2px}{z} = 1.00. \tag{8}$$

The nature of this relationship makes it difficult to solve analytically for x and p, given a specific value of ρ. However, the comprehensive normal curve tables of Kelley [5] permit a sufficiently exact approximation for p for select values of ρ. A number of such pairs of values are tabulated in Table 1.

These values indicate that the definition of the optimal extreme groups remains remarkably constant over a rather wide range of values for ρ. For $\rho = .10$, the function reaches a maximum at $p = .27$; for $\rho = .80$, the maximum occurs at $p = .23$. This finding has important implications for researchers employing the difference approach for testing relationships. It suggests that extreme groups of from 25 to 27 percent provide the most powerful test of the existence of a moderate linear relationship. This size group is especially appropriate when the relationship is weak, as it usually is in the experimental context here considered.

From a practical point of view, it is extremely fortunate that this maximum power is reached when groups smaller than the upper and lower halves are employed. Frequently the nature of the experimental treatment or the need to share the pool of experimental subjects forces the investigator to

TABLE 1

Extreme Groups which Result in Mean Difference Tests of Maximum Power

ρ	Percent in Each Extreme Group
.10	27.0
.20	26.9
.30	26.6
.40	26.3
.50	25.8
.60	25.1
.70	24.2
.80	23.3

use only a fraction of those originally tested. These results prove that such a restriction can result in greater, rather than less, power in the crucial statistical tests.

Comparison of Difference and Correlation Approaches

If a limited subgroup of subjects may be used to investigate the presence of a linear relationship between X and Y, should the experimenter draw all of his subjects from the extreme portions of the X distribution and test the difference between Y means, or should he draw subjects at random from the entire range on X, estimate the linear correlation, and test it for significance? The answer to this question is clearly pertinent to the development of an adequate research strategy. As in the previous discussion, design efficiency will be evaluated by the power of the statistical tests that are involved.

The test of significance of a product-moment correlation or, more precisely, the test of the linear regression coefficient is

$$t = \frac{r\sqrt{N_0 - 2}}{\sqrt{1 - r^2}}. \tag{9}$$

The power parameter for this test may be obtained by substitution of the appropriate values into formula (1). In this case, the regression coefficient is the normally distributed variable; the population standard deviation is the standard error of the regression coefficient, $\sigma_Y \sqrt{1 - \rho^2} / \sqrt{N_0}\, \sigma_X$. After algebraic simplification, ϕ is found to equal

$$\phi_r = \frac{\beta}{\sqrt{2}\sigma_b} = \frac{\rho\sigma_Y/\sigma_X}{\sqrt{2}\sigma_Y\sqrt{1 - \rho^2}/(\sqrt{N_0}\sigma_X)} = \frac{\rho\sqrt{N_0}}{\sqrt{2}\sqrt{1 - \rho^2}}. \tag{10}$$

With $2pN$ subjects this reduces to

$$\phi_r = \frac{\rho\sqrt{pN}}{\sqrt{1-\rho^2}}. \tag{11}$$

For the difference test based on pN subjects per group, the value of ϕ reduces to

$$\phi_d = \frac{z\rho}{\sqrt{\frac{p(1-c\rho^2)}{N}}}. \tag{12}$$

The comparative power of the difference test and the correlation test may be most easily inferred from the ratio of ϕ_d to ϕ_r. Where the ratio exceeds 1.0, the difference test is more powerful; where the ratio is less than 1.0, the correlation test is more powerful.

In general terms, the ratio equals

$$\frac{\phi_d}{\phi_r} = \frac{z/\sqrt{p(1-c\rho^2)}}{\sqrt{p}/\sqrt{1-\rho^2}}. \tag{13}$$

It may be noted that the value of the ratio depends upon both p and ρ. The magnitude of ρ is, of course, unknown. The value of p, on the other hand, is partially within the control of the experimenter and partially dictated by the nature of the experiment. In some situations the experimenter might be limited by practical considerations to the use of no more than a small proportion of the subjects in the pool. In other instances, it might be quite feasible to obtain X and Y measures on all N subjects.

To reveal the conditions under which each approach is the more powerful, the ratio was evaluated for various conditions of subject availability. These conditions, which indicate the proportion of subjects available for the second stage of the experiment, range from 20 to 100 percent. (Since the difference technique achieves close to maximum power when upper and lower quarters are used, the value of ϕ_d was based on $p = .25$ for all availability percentages above 50.) For each experimental condition the value of ρ was determined which would make the ratio greater than 1.0. These results are reported in Table 2.

The data in this table reveal that as the availability percentage increases, the choice of design strategy gradually shifts in favor of the correlation approach. However, the advantage of the extreme group design holds until the availability percentage is quite large or the population correlation is quite high. The two-stage design here considered would generally be applied in instances where only a moderate degree of relationship holds. Values of .50 or higher are probably relatively rare in this preliminary stage of construct definition. In view of this fact, a fairly clear-cut recommendation may be made. If less than 75 percent of the subject pool can be used in the

TABLE 2

Comparative Power of Difference and Correlation Approaches for Testing the Presence of Relationship

Percent of Subjects Available for Treatments	More Powerful Approach
20	Difference when $\rho < .962$ Correlation when $\rho > .962$
30	Difference when $\rho < .938$ Correlation when $\rho > .938$
40	Difference when $\rho < .902$ Correlation when $\rho > .902$
50	Difference when $\rho < .847$ Correlation when $\rho > .847$
60	Difference when $\rho < .768$ Correlation when $\rho > .768$
70	Difference when $\rho < .624$ Correlation when $\rho > .624$
75	Difference when $\rho < .492$ Correlation when $\rho > .492$
80	Difference when $\rho < .198$ Correlation when $\rho > .198$
More Than 80	Correlation for all values of ρ

experiment, the difference approach will almost surely be the more powerful. If more than 80 percent of the subjects can be employed in the second portion of the experiment, the correlation approach will almost certainly be the more powerful.

The above comparison does not take into account the added degrees of freedom of the test of r when the availability percentage is greater than 50. Therefore, the values of ρ noted in Table 2 must be regarded as approximate. However, the power function charts of Pearson and Hartley indicate that for a given value of ϕ the power of a t test is only slightly affected by increases in the degrees of freedom beyond 60. Thus, the recommendation above seems sufficiently precise for all practical purposes.

The difference in power of the two approaches may be illustrated by an example. Assume $\rho = .30$ and that a total of 100 subjects was initially tested on the classification variable. If the experimenter used upper and lower

quarters, he would have a t test with 48 degrees of freedom. With a 5 percent level of significance, the power of this test would be approximately .78. If, on the other hand, 50 subjects were selected from the entire range of the classification variable and the product-moment coefficient were tested for significance at the same level, the power would equal .57. If 75 subjects were selected from the full range on X, the power of the correlation test would equal .75. If all 100 subjects could be used, the power would equal .87.

At first glance, it may appear paradoxical that throwing away data from the middle half of the distribution improves the difference approach and renders it superior to correlation analysis based on a greater number of subjects. However, these results are consistent with the findings of other investigators [1, 3]. Those familiar with common item analysis procedures will no doubt recognize the analogous results which hold in that field. Indeed, Kelley's proof [6] that item discrimination is most efficiently assessed through the use of the upper and lower twenty-seven percents follows a very similar line. The discard of the middle portion of the distribution does not merely reduce the *quantity* of the information—a practice which rarely, if ever, works to benefit of a statistical procedure—but also changes the *nature* of data. When the full import of the modifications of both quantity and quality are appreciated, the result no longer seems quite so anomalous.

It should be emphasized that these results apply only to data which conform to the bivariate normal distribution. Of particular importance is the assumption of linear correlation. Where the hypothesis of a curvilinear relationship can be seriously entertained, it would clearly be unwise to sample in a fashion which did not permit close study of the nature of the relationship. Thus preference should be given to the difference approach only when the assumption of linearity is strongly tenable.

Estimating the Correlation from Extreme Group Statistics

The foregoing development has been concerned only with the evaluation of the presence of a linear relationship, not with estimation of the strength of that relationship. As McNemar [8] has succinctly pointed out, such a methodology is almost certain to be abused, for it can easily lead the experimenter to exaggerate the importance of trivial results. If the distributions of X and Y are normal, or approximately so, and if the relationship is linear, a useful approximation of the product-moment coefficient may be obtained from statistics derived from the extreme groups.

From (6) the difference between means is seen to equal

$$\mu_{Y_U} - \mu_{Y_L} = \frac{2\rho z \sigma_Y}{p},$$

and the variance within either extreme group, as derived in [2], equals

$$\sigma^2_{Y_U} = \sigma^2_Y\left(1 - \rho^2\left[\frac{z^2}{p^2} - \frac{xz}{p}\right]\right).$$

Thus

$$\sigma_Y = \frac{\sigma_{Y_U}}{\sqrt{1 - \rho^2\left[\frac{z^2}{p^2} - \frac{xz}{p}\right]}},$$

and

$$\mu_{Y_U} - \mu_{Y_L} = \frac{2\rho z\sigma_{Y_U}}{p\sqrt{1 - \rho^2\left[\frac{z^2}{p^2} - \frac{xz}{p}\right]}}.$$

Solving this equation for ρ yields

$$\rho = \frac{\mu_{Y_U} - \mu_{Y_L}}{\sqrt{\frac{4z^2\sigma^2_{Y_U}}{p^2} + \left(\frac{z^2}{p^2} - \frac{xz}{n}\right)(\mu_{Y_U} - \mu_{Y_L})^2}}. \tag{14}$$

If the upper and lower quarters are used to test for the presence of a relationship, (14) becomes

$$\rho = \frac{\mu_{Y_U} - \mu_{Y_L}}{\sqrt{6.4630\sigma^2_{Y_U} + (.7584)(\mu_{Y_U} - \mu_{Y_L})^2}}.$$

Using sample means to estimate population means and the mean square within groups to estimate the variance within extreme populations, the estimate of ρ becomes

$$\tilde{\rho} = \frac{\bar{Y}_U - \bar{Y}_L}{\sqrt{6.4630\ \mathrm{MS}_{\mathrm{within}} + .7584(\bar{Y}_U - \bar{Y}_L)^2}}. \tag{15}$$

This estimate represents a solution to a special case of the problem of estimating ρ from data obtained on extreme groups. A solution for the more general problem, in which no restriction is imposed on the comparative size of the extreme groups, has been presented by Peters and Van Voorhis [9]. Such estimates should be used, of course, only when the assumptions of normality and linearity are tenable.

REFERENCES

[1] Bartlett, M. S. Fitting a straight line when both variables are subject to error. *Biometrics*, 1949, **5**, 207-212.

[2] Feldt, L. S. A comparison of the precision of three experimental designs employing a concomitant variable. *Psychometrika*, 1958, **23**, 335-353.

[3] Gibson, W. M. and Jowett, G. H. "Three-group" regression analysis. I. Simple regression analysis. *Appl. Statist.*, 1957, **6**, 114-122.

[4] Kelley, T. L. *Fundamentals of statistics.* Cambridge, Mass.: Harvard Univ. Press, 1947.
[5] Kelley, T. L. *The Kelley statistical tables.* Cambridge, Mass.: Harvard Univ. Press, 1948.
[6] Kelley, T. L. The selection of upper and lower groups for the validation of test items. *J. educ. Psychol.*, 1939, **30**, 17-24.
[7] McClelland, D. C. et al. *The achievement motive.* New York: Appleton-Century-Crofts, 1953.
[8] McNemar, Q. At random: Sense and nonsense. *Amer. Psychologist*, 1960, **15**, 295-300.
[9] Peters, C. C. and Van Voorhis, W. R. *Statistical procedures and their mathematical bases.* New York: McGraw-Hill, 1940.
[10] Pearson, E. S. and Hartley, H. O. *Biometrika tables for statisticians, Vol. I.* Cambridge, England: University Press, 1956.
[11] Taylor, J. A. Drive theory and manifest anxiety. *Psychol. Bull.*, 1956, **53**, 303-320.

Manuscript received 5/7/60

Revised manuscript received 12/5/60

PSYCHOMETRIKA—VOL. 26, NO. 3
SEPTEMBER, 1961

THE RATIONALE FOR AN "OBLIMAX" METHOD OF TRANSFORMATION IN FACTOR ANALYSIS*

D. R. SAUNDERS†
PRINCETON, NEW JERSEY

Factorial transformation is viewed as an estimation problem in which the usual assumption of homogeneously distributed error cannot be applied, but may be replaced by a principle of maximum kurtosis. This leads to quartimax in the orthogonal case, and to "oblimax" in the oblique case. Oblimax is readily programmable, and typically provides results similar to those of subjective rotation. However, oblimax may encounter special difficulty in data which do not determine a simple structure, or which have been imprecisely factored.

Four independent lines of investigation converged in 1953 on the "quartimax" criterion for fitting an orthogonal simple structure to a sample of vectors in k-space. Carroll's original solution [1] was obtained by minimizing the function

$$f = \sum_i a_{im}^2 a_{in}^2 \tag{1}$$

simultaneously for every pair of factors, m and n, where a_{im} is the correlation of vector i with factor m. Saunders [8] obtained an equivalent solution by maximizing the function

$$K = \sum_i \sum_j a_{ij}^4 / (\sum_i \sum_j a_{ij}^2)^2. \tag{2}$$

The rationale of minimizing f, or of maximizing K, was based on Thurstone's rules for a determinate simple structure in factor analysis [9]. Neuhaus and Wrigley [5] also obtained an equivalent solution by maximizing the variance of squared factor loadings, i.e., by maximizing the function

$$V = \sum_i \sum_j (a_{ij}^2 - \overline{a_{ij}^2})^2. \tag{3}$$

*The Managing Editor has substituted the word "transformation" for the word "rotation" in the title and throughout this paper, on the grounds that "oblique rotation" is a self-contradictory term, the use of which need not be perpetuated.

†This paper is primarily a condensation of material that first appeared in ETS Research Bulletins 53–10 and 54–31, both long out of print. In this treatment the principle of maximum kurtosis receives increased emphasis, and the special case for equation (10) is recognized. The writer is indebted to his former colleagues, Mr. Charles Pinzka and Dr. Ledyard Tucker, for invaluable assistance in achieving a straightforward and general derivation of equation (10).

At about the same time, Ferguson [3], commencing from a fresh interpretation of parsimony in factor analysis, arrived at the suggestion that

$$Q = \sum_i \sum_j a_{ij}^4 \tag{4}$$

be maximized as a basis for factorial rotation. The equivalence of these four formulations for orthogonal rotation can easily be demonstrated, since each of the first three then reduces to the fourth. The term "quartimax," suggested by Burt, can be applied equally to any of them.

The equivalence of these formulations results from the constancy of $\sum_i \sum_j a_{ij}^2$ under orthogonal rotation. Since this quantity may vary in oblique transformation it is evident that the four formulations are no longer interchangeable, and the fact that they are numerically distinct may in principle be shown by simple examples using only two factors and three or four variables. Under these circumstances it appears worthwhile to consider the relative merits of deriving extensions of these methods to include the oblique case.

Applicability of a Principle of Maximum Kurtosis

We may observe at once that the simple maximization of Q does not lend itself to oblique extension, since it will be difficult to insure that exactly k distinct factors will be obtained from the transformation. Carroll himself reported the extension of formulation (1) to the oblique situation [1], but this approach has suffered from the use of a definition involving the relation of two factors considered together. For one thing, this led originally to matrices of order $k(k-1)$, which would be unmanageable for moderate values of k even on a computer; Kaiser [4] has since provided a solution using matrices of order k. At a still more fundamental level, it seems to us that it would be desirable to be in a position to evaluate the "quality" of the simple structure for any one given factor independently of the company of other factors in which it happens to be found. Formula (1) does not lend itself to such use so well as (2) or (3).

The notion of theoretically *evaluating* the quality of a simple structure one factor at a time will not necessarily alter the fact that this quality is *determined* in practical analysis—especially in orthogonal rotation—by the relation of the factor to other factors according to our present imperfect concepts of simple structure. Less directly, the selection of variables, the choice of communalities, and the completeness of factor extraction will also contribute their influence to this determination. Our wish would be to evaluate the aggregate of all such effects on a *particular* factor.

A rational basis for a choice between formulas (2) and (3) is more difficult to find, but one may perhaps be found by regarding factorial transformation as a problem in statistical estimation. One traditional technique

for estimation is to apply the familiar principle of least squares. Application of this principle to the rotation problem leads directly to the principal axes, commencing with the smallest one, and these are properly recognized as inferior to simple structure axes for many purposes. The principle of least squares requires an assumption of homogeneously distributed error that simply cannot be satisfied in the factor-analysis situation. If we assume that all nonzero factor loadings reflect error, as this model requires, then we may never interpret a high loading in terms of its psychological or other meaning. If we wish to interpret high loadings, we must reject the principle of least squares as a guide to rotation.

A more general technique that may be applied to estimation problems is the principle of maximum likelihood. Despite its generality, however, this principle is impotent except in the presence of a model spelling out the sampling process that generates the distribution of measurement error. Such a model for the error of a factor loading has never been adequately worked out. In practice, the principle of maximum likelihood usually reduces to a least squares solution, and it seems probable that there are a limited number of coordinate alternatives to least squares for us to consider.

One possible alternative to least squares is a "principle of maximum kurtosis." Although such a principle does not appear to have been previously stated, it seems likely to apply to a wide variety of everyday estimation problems provided one is prepared to cope with relatively complicated numerical computations. Even in the simplest application of estimating the location parameter of a distribution a cubic equation must be solved, and either one or two roots of the equation may be meaningful. (Two are meaningful when the basic distribution is U-shaped.) In order to fit a straight regression line, a pair of simultaneous cubic equations must be solved.

The principle of maximum kurtosis is based directly on an assumption that deviations from the measure of central tendency are *not* homogeneously distributed. Application of the principle may be appropriate whenever some deviations are expected to scatter more widely than others—either as a result of special or freak errors affecting isolated observations or, as in factor analysis, where meaningful effects are supposed to produce occasional big deviations that are independent of the error variance.

Formulation (2) may be regarded as a direct application of the principle of maximum kurtosis to the rotation problem, and for this reason it seems to afford more justification for extension into the oblique case even than formulation (3). Application of the principle is made comparatively easy because the expected value for a factor loading is given by theory as zero, and ordinarily no measure of central tendency will have to be estimated from the data. It is a main purpose of this paper, therefore, to investigate the implications of this rationale. The computing procedure that will be implied has actually been in use for several years under the name of "oblimax,"

and it is a corollary purpose of this paper to provide a theoretical foundation for what appears to be a practically valuable computational technique.

It should be noted that there *are* factor-analysis situations in which artificial contraints are imposed on a set of factor loadings, such as a zero-sum property when the scores were rank-order data. In such cases we must expect to combine translation with rotation in the search for simple structure.

Derivation of the Oblimax Criterion

Let a_i be the desired loading of variable i on factor A after rotation. We seek to find positions for reference vectors that will make

$$K = \sum_i a_i^4 / (\sum_i a_i^2)^2 \tag{5}$$

a relative maximum with respect to rotation in any plane defined by A in combination with any other reference vector. Let b_i and c_i be the loadings of variable i on two arbitrary vectors, B and C, that define a plane containing A. B and C need not be orthogonal. Then

$$K = \frac{\sum_i (b_i u_b + c_i u_c)^4}{[\sum_i (b_i u_b + c_i u_c)^2]^2}, \tag{6}$$

where u_b and u_c are weights that may be used to obtain A from B and C. If $u_b \neq 0$, we may define x as equal to u_c/u_b and write

$$K = \frac{\sum_i (b_i + c_i x)^4}{[\sum_i (b_i + c_i x)^2]^2} \equiv \frac{R}{S^2}. \tag{7}$$

Using primes to designate differentiation with respect to x, a necessary condition for maxima is that

$$K' = \frac{S^2 R' - 2RSS'}{S^4} = 0, \tag{8}$$

which leads to

$$R'S - 2RS' = 0. \tag{9}$$

Resubstitution of the values of R and S and of their derivatives with respect to x produces a collection of polynomial terms in x, with coefficients formed from various sums of powers and cross products of the b_i and c_i. All terms higher than the fourth degree in x conveniently cancel, and we are left with

$$
\begin{aligned}
&\Big(\qquad\qquad\qquad \sum bc \sum c^4 \; - \; \sum c^2 \sum bc^3 \Big) x^4 \\
&+ \Big(\sum b^2 \sum c^4 \; + 2 \sum bc \sum bc^3 - 3 \sum c^2 \sum b^2c^2 \Big) x^3 \\
&+ \Big(3 \sum b^2 \sum bc^3 \qquad\qquad - 3 \sum c^2 \sum b^3c \Big) x^2 \\
&+ \Big(3 \sum b^2 \sum b^2c^2 - 2 \sum bc \sum b^3c - \sum c^2 \sum b^4 \Big) x \\
&+ \Big(\sum b^2 \sum b^3c \; - \; \sum bc \sum b^4 \qquad\qquad \Big) = 0.
\end{aligned} \tag{10}
$$

Unless vectors B and C are collinear, in which case they did not *define* a plane in the first place, (10) should have nonzero coefficients and four roots. In the typical case these roots will all be real and distinct, and will correspond to two relative maxima and two relative minima for K. In addition, we shall have to consider (in the next section) a special case with only one maximum and one minimum for K, and two unusable complex roots from (10). It is also possible to construct singular cases with multiple roots and indeterminate cases in which all the coefficients of (10) vanish despite the use of distinct vectors for B and C, but the probability is zero that such singular or indeterminate cases will ever arise out of data subject to measurement error.

In the typical case we may distinguish the roots that correspond to maxima for K by examining the second derivative, which should be negative. It turns out that

$$K'' = \frac{S^3(R'S - 2RS')' - (R'S - 2RS')3S^2S'}{S^6}, \tag{11}$$

or, using (9),

$$K'' = (R'S - 2RS')'/S^3. \tag{12}$$

S is a sum of squares and must always be positive. Therefore, the sign of K'' is identical with the sign of the derivative of the polynomial equation. As x increases without limit, the sign of the x^4 term will determine the sign of both the polynomial and its derivative. Therefore, if the sign of x^4 is negative in (10), the algebraically largest root will yield a maximum for K, and if the sign of x^4 is positive it will yield a minimum for K. Since an algebraic ordering of the roots must provide for alternation between maxima and minima, the identity of the remaining roots is revealed.

An important feature of the solution that is offered by the typical roots of (10) is that two distinct roots are provided, which we may christen as x_{A_1} and x_{A_2}, each of which satisfies the requirements for a relative maximum for K. While we have no basis for choosing one of these as "better" than the other, neither is it necessary for us to do so. A highly satisfactory solution may be obtained by considering successive pairs of factors, in each case substituting *two* new factors for *both* of the old factors. Since the two new

factors are distinct, they continue to define the plane in which they were determined, and the entire set of k factors continues to span the k dimensions with which the process began. It is not necessary to establish any arbitrary limitation on the closeness that two transformed factors may exhibit to each other.

(A *meaningful* limitation on the obliquenesses of factors may be conceived in terms of a restriction on the rank of the matrix of factor intercorrelations. By limiting the number of second-order factors to one or a few, the advantages of oblique transformation would be preserved without the wastage and risk of indeterminacy implicit in estimating large numbers of supposedly independent transformation cosines on the basis of the degrees of freedom afforded by the data. It would be helpful to know how such a restriction on rank could be enforced.)

On the other hand, in planning for automatic iterations of the method on a digital computer, it will be helpful to be sure that each new factor is used to replace the old factor which it most resembles. A major fraction of the computer's time will be spent in searching for roots for (10). Some of this time may be saved by noting that, as the axes move toward their final positions, one of the desired roots and the reciprocal of the other one will always be very close to zero.

(While it is possible to adopt a formula solution for the roots of a quartic equation (e.g. [7], p. 154), the method is cumbersome at best and appears to require needlessly many instructions in relation to the memory capacity of such a machine as the IBM 650. On the other hand, an exclusive dependence on Newtonian approximation via synthetic division fails to provide for the special case. An optimum strategy appears to be (i) find any two of the real roots by successive approximation, (ii) divide them out, and (iii) solve the resulting quadratic equation by formula.)

In order to obtain the desired loadings on vectors A_1 and A_2 to complete a single iteration, we must first recover u_b and u_c for each of the roots, x_{A_1} and x_{A_2} . This may be effectively accomplished by the procedure usually employed in oblique single-plane graphical solution, since x_{A_1} and x_{A_2} are simply the tangents of the indicated transformations from B towards C as they would appear in an orthogonalized factor plot. As a practical procedure we form the matrix product

$$\Lambda_{BC}\begin{bmatrix} 1 & 1 \\ x_{A_1} & x_{A_2} \end{bmatrix} = \Lambda^*_{A_1A_2} , \tag{13}$$

where Λ_{BC} is the two-column transformation matrix required to produce B and C from the original orthogonal basis for rotation. (Initially, these are two columns from an identity matrix.) We then find $\Lambda_{A_1A_2}$ by normalizing the columns of Λ^*, to make the sums of squares unity. A_1 and A_2 are found

by applying Λ_{A,A_s} to the original basis. By this procedure, the Λ relating the final output to the original input must be obtained as a by-product of the computation.

Computer programs for performing "oblimax" transformation according to this general procedure have been written for Illiac [2], the basic IBM 650 [6], the augmented IBM 650,* and possibly for other machines. These programs vary in their ability to make concurrent internal use of the device of vector normalization, which may be as advantageous with oblimax as it is with quartimax. None of these programs appear to deal fully with the "special case." A program for the IBM 7090 is in preparation.

Explanation of the Special Case

Configurations of data that contain intuitively recognizable evidence of good simple structure do lead into the typical case considered in the last section. However, it is possible to distribute the vectors of the variables in the *BC* plane in such a way that (10) has only two real roots, with the other two complex. For example, the following set of seven points will produce such a "special" result: (0, 2), (2, 0), (1, 1), (1, 1), (1, 1), (1, 1), (1, 1). The real roots in this example are $x_1 = -1$ and $x_2 = 1$. The complex roots satisfy the quadratic equation $5x^2 + 7x + 5 = 0$. Consideration of the factor plot provided by these seven vectors suggests that x_1 and x_2 actually do define an appropriate rotation. Although x_2 now corresponds to a minimum for K, we may regard it as defining the reference vector for a residual (nonmeaningful) factor whose hyperplane is for some reason displaced from zero. (This may occur, for example, as a result of poor communality estimates.) Conveniently, the single meaningful factor that appears to be present in this situation is made orthogonal to the residual factor.

In practice, such special configurations have turned out to be of appreciable importance. All of the examples we have seen conform to the "displaced residual" paradigm just illustrated. While these configurations are seen most often in the early stages of transformation by this method, they may persist in an otherwise fully convergent structure. In the event of a truly nonmeaningful factor, we should expect this type of configuration to characterize most of its factor plots with other factors; such factors may be eliminated from interpretation altogether. On the other hand, if this special situation persists in just one or a few isolated factor plots, it may have to be conceded that the basic data simply are inadequate to define a determinate oblique simple structure; we can neither disregard the factors nor be sure how to interpret them. It would indeed be surprising if the latter situation never materialized.

*According to Dr. Steven Vandenberg, this program was never made fully operational because the Nickles-Keenan program [6] became available to do the same thing with less equipment.

REFERENCES

[1] Carroll, J. B. An analytical solution for approximating simple structure in factor analysis. *Psychometrika*, 1953, **18**, 23-38.
[2] Dickman, K. W. KSL 1.90 Oblimax. *Behav. Sci.*, 1959, **4**, 339. (Abstract)
[3] Ferguson, G. A. The concept of parsimony in factor analysis. *Psychometrika*, 1954, **19**, 281-290.
[4] Kaiser, H. F. Note on Carroll's analytic simple structure. *Psychometrika*, 1956, **21**, 89-92.
[5] Neuhaus, J. O. and Wrigley, C. F. The quartimax method: An analytical approach to orthogonal simple structure. *Brit. J. statist. Psychol.*, 1954, **7**, 81-91.
[6] Nickles, M. R. and Keenan, T. A. Minimal IBM 650 program for the oblimax rotation to simple structure. *Behav. Sci.*, 1960, **5**, 100. (Abstract) IBM 650 Program Library: File Number 6.0.505.
[7] Rietz, H. L. and Crathorne, A. R. *College algebra.* (3rd ed.) New York: Holt, 1937.
[8] Saunders, D. R. An analytical method for rotation to orthogonal simple structure. *Amer. Psychologist*, 1953, **8**, 428. (Abstract)
[9] Thurstone, L. L. *Multiple-factor analysis.* Chicago: Univ. Chicago Press, 1947.

Manuscript received 5/3/60

Revised manuscript received 2/28/61

PSYCHOMETRIKA—VOL. 26, NO. 3
SEPTEMBER, 1961

MULTIDIMENSIONAL UNFOLDING: SOME GEOMETRICAL SOLUTIONS

D. W. McElwain and J. A. Keats
UNIVERSITY OF QUEENSLAND

A method is proposed whereby the distribution of stimuli and subjects can be derived directly from the subjects' rankings of the stimuli. The solution for four stimuli is presented with examples and suggestions for the solution of problems involving more stimuli. The model is intuitively attractive and is likely to prove useful when direct solutions for larger number of stimuli are available.

At the time of publication of Bennett and Hay's article [1] on multidimensional unfolding, the senior author was analyzing data in terms of the model presented in this paper. The power of the model in representing both persons and stimuli in the same space makes it attractive for use with certain types of social data. A simple solution is proposed to the problem with four stimuli, together with some suggestions for the solution of more complex problems. Considerations which should be borne in mind when this model is being used will also be noted.

The Problem

In the course of a study of the leisure time activities of school children, the children were asked to list in order of preference the radio stations to which they listened. In nearly all cases four radio stations in the Brisbane area were mentioned, but only 18 of the possible 24 rank orders were observed with 304 subjects. As noted by Hays [1], this number of rank orders can be generated by four stimuli in two dimensions using the unfolding principle. The problem that arises is that of discovering whether there is a configuration of the stimuli in two dimensions which could generate this particular set of 18 rankings, and also discovering the properties of such a configuration if it exists. The solution proposed is geometrical so that both aspects of the problem are solved. In this way it is superior to those suggested by Hays which only consider the first aspect. The data obtained from the ordering of radio stations is presented in Table 1. The possible orders are grouped according to the stimuli placed last. A solution in one dimension accounts for only about 70 percent of the cases so that the existence of two or more dimensions is clearly indicated.

TABLE 1

Children's Preferences for Four Radio Stations

ABCD	112	ABDC	25	ACDB	12	BCDA	-
ACBD	59	ADBC	13	ADCB	8	BDCA	-
CABD	10	DABC	4	CADB	5	CBDA	2
BACD	25	BADC	6	DACB	4	DBCA	-
CBAD	7	DBAC	1	CDAB	-	CDBA	-
BCAD	3	BDAC	-	DCAB	7	DCBA	1

The Solution

The possible geometrical solutions can be classified in terms of the number of times each stimulus is placed last, i.e., (6, 6, 6, 0) is the solution in which three stimuli are placed last in all possible ways and the fourth is never placed last. In this notation a stimulus may be designated by the number of times it occurs in last position, e.g., the stimulus (0). In a similar way the six orders (0, 6, 6, 6) refer to the group of orders in which the stimulus (0) appears first in all possible ways; or again, the triangle (6, 6, 6) refers to that triangle formed from the locations of the three points each of which appears last in all possible ways. This notation permits a more concise discussion of the various solutions. It is possible in practice to obtain a particular solution type which does not fulfill the additional conditions. It may however still be possible to choose a solution type involving one or more additional orders for which the sufficiency conditions are fulfilled, e.g., a (2, 2, 2, 2) case may not fit the conditions for a square, but may fit those for a cyclic trapezium, with, of course, some orders not occurring. Table 2 shows the possible solutions.

A note on concave solutions. In general if a stimulus is never ranked last, then this stimulus is located inside the closed polygon with the remaining stimuli as vertices. In other words the polygon with all stimuli as vertices is concave. In such cases the problem immediately reduces to one of $S - 1$ stimuli in at most $S - 2$ dimensions, where S is the number of stimuli. There will of course be S sets of $S - 1$ stimuli but it may not be necessary to consider all of these. Thus the general problem depends *solely* on convex polygons. It seems possible that general solutions can be derived by considering slight deviations from cyclic polygons.

Outlines of proofs of the properties stated in Table 2 are given in the Appendix.

Application of the method

An example from Bennett and Hays. In their recent article ([1], p. 36) Bennett and Hays take as an illustrative example the following set of 18

TABLE 2
Solutions of the Four Stimuli Unfolding Problem

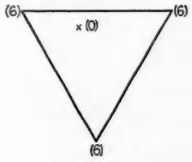
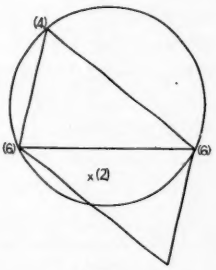
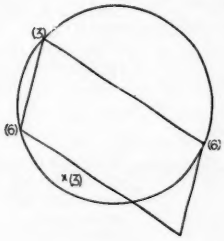
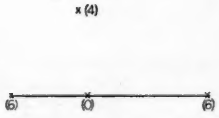
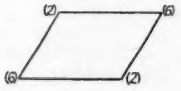
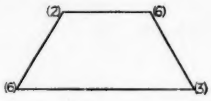
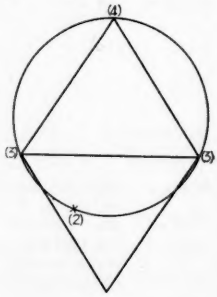
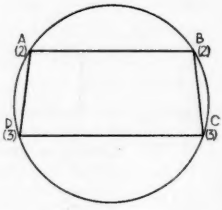
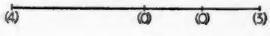
Solution Type	Sufficient Conditions	Geometrical Properties	Diagram
(6,6,6,6)	Nil	Any non-degenerate tetrahedron	
(6,6,6,0)	Nil	The point (0) is inside the triangle (6,6,6). This is the only case for which the quadrilateral is concave.	
(6,6,4,2)	The two orders ending in (2) are (4,6,6,2) and the two missing orders ending in (4) are (6,6,2,4)	The point (2) lies inside the circle (6,6,4) and the Parallelogram formed by 6,6, and 4 with 6,6 as diagonal, but outside the triangle (6,6,4). (4) lies inside the circle (6,6,2).	
(6,6,3,3)	The four orders (6,6,3,3) are missing as are two of the four orders (6,3,6,3) beginning with the same point (6)	Each point (3) lies inside the circle (6,6,3) but outside the triangle (6,6,3) and the parallelogram formed by 6,6 and 3. (3) and (3) are on opposite sides of (6) and (6). The line (3,3) is closer to the point (6) which begins the two missing orders (6,3,6,3).	
(6,6,4,0)	The two orders (6,6,0,4) are missing as well as all orders ending in (0).	The points 6,0,6 are collinear in that order. The point (4) is not in that line or its extension.	
(6,6,2,2)	The four orders (2,6,6,2) occur, but no others ending in (2).	Parallelogram with (6), (6) as major diagonal.	

TABLE 2—Continued

Solution Type	Sufficient Conditions	Geometrical Properties	Diagram
(6,6,3,2)	The four orders (2,6,6,3) and (3,6,6,2) occur with one order (6,2,6,3)	Trapezium, with the shorter diagonal, (2), (3) further from the point (6) appearing first in the order (6,2,6,3) which does occur.	
(4,3,3,2)	The two orders ending in (2) are (4,3,3,2) and the two orders missing from (4) are (3,3,2,4). The six orders (3,2,4,3) (3,4,2,3) and (4,3,2,3) occur.	Cyclic quadrilateral with diagonals (4), (2) and (3), (3). (2) is inside the parallelogram defined by (3), (4) and (3).	
(3,3,2,2)	With the lettering in the diagram, the orders DABC, ADBC, BACD and BCAD and their opposite occur.	The line (2), (2) and the line (3), (3) form the shorter and longer parallel sides of a cyclic trapezium. Points (2) and (3) which never occur together at the beginning or ending of an order form diagonals.	
(2,2,2,2)	With the lettering of the diagram the eight orders are ABDC, ADBC, BACD and BCAD with their opposite.	Rectangle or square. Diagonal points never appear together at the beginning or ending of orders.	
(4,3,0,0)	One order (4,0,0,3) and its opposite with consequent five orders.	Straight line in the order (4,0,0,3) with segment (4), (0) greater than segment (0), (3).	
(3,3,0,0)	The two orders (3,0,0,3) with consequent five orders.	Straight line in the order (3,0,0,3) with segment (3), (0) equal to segment (0), (3).	

of the possible 24 orderings of four stimuli. This set has been rearranged so that orders placing a particular stimulus last are grouped in columns.

ABCD	CDAB	CBDA	ADBC
ACBD	ACDB	CDBA	ABDC
BCAD	ADCB	DCBA	DABC
CBAD	DCAB		
CABD	CADB		
BACD	DACB		

This grouping reveals a possible (6, 6, 3, 3) solution. Four of the missing orders are (6, 6, 3, 3) orders and the remaining two are (6, 3, 6, 3) orders both beginning with *B*, viz., BCDA and BADC. From the table of solutions it can be seen that the necessary and sufficient conditions are fulfilled for these orders to be represented by four points in two dimensions. In particular, B is closer to the line AC than D as shown in ([1], p. 41). All four stimuli problems can be solved in this way for two or one dimensions.

Application to the problem of the present paper. In considering the problem that initiated this study, it is immediately clear that this would require a (6, 5, 5, 2) solution, but there is no such solution. However, it is of some interest to see how well the two-dimensional model does fit these data. Reference to the table of solutions shows that there is a (6, 6, 4, 2) solution with only two discrepancies in 304 cases, a (6, 6, 6, 0) solution with three discrepancies, a (6, 6, 4, 0) solution with four discrepancies, a (6, 6, 3, 3) solution with six discrepancies, and an alternative (6, 6, 3, 3) solution with seven discrepancies. The next closest fitting solution involves ten discrepancies, seven of which are for the same ordering and this and other alternatives may safely be ignored. The solutions are displayed in Figs. 1–5. Figs. 1, 2, and 3 stress the central position of the most popular

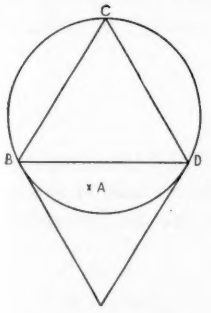


FIGURE 1
(6,6,4,2) Solution
(Two Discrepancies)

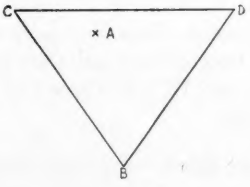


FIGURE 2
(6,6,6,0) Solution
(Three Discrepancies)

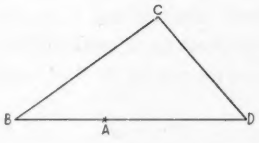


FIGURE 3
(6,6,4,0) Solution
(Four Discrepancies)

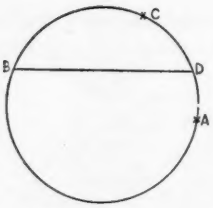


FIGURE 4
(6,6,3,3) Solution
(Six Discrepancies)

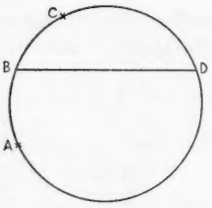


FIGURE 5
(6,6,3,3) Solution
(Seven Discrepancies)

station, A. Figures 1, 4, and 5 stress the contrast of B and D, with A and C contrasted on a different dimension and in a certain sense closer to each other. Both of these characteristics are shown in Fig. 1 with neither being unduly emphasized; this gives the best fit (less than one percent of cases deviating). It is worth noting that D is the government radio station which is free from advertising whereas A, B, and C are commercial stations with B concentrating a considerable amount of its time on short advertisements. Station A has a large number of imported serials in its children's session, and these appeal to the children of the age group studied.

If the cardinality and groups criteria of Bennett and Hays [1] are applied to this problem they both indicate a minimum dimensionality of two. However, if the more sensitive criterion that for four stimuli in two dimensions there must not be more than four pairs of regions differing only in the order of two stimuli ([1], p. 43) is applied, two dimensions are indicated for five of the six possible pairs, but the sixth pair A, B gives the following results.

ABCD	(112)—	BACD	(25)
CABD	(10)—	CBAD	(7)
ABDC	(25)—	BADC	(6)
DABC	(4)—	DBAC	(1)
DCAB	(7)—	DCBA	(1)
CDAB	(–)—	CDBA	(–)

The numbers indicate frequencies. Thus five pairs occur and these require three dimensions. However, two of these pairs are present because of single occurrences of orders DBAC and DCBA. These two cases are the discrepancies for the (6, 6, 4, 2) solution.

Some Questions of Methodology

One of the difficulties presented by the unfolding model is that with reasonable numbers of stimuli and dimensions the number of possible rank orders is very large. However, it might also be argued that there exist only

a small number of stimuli of a given class about which subjects feel strongly enough to restrict its orders of preference to a number well below the maximum. This is not a technique which is likely to be useful if the number of dimensions is artificially increased, e.g., by including preferences for political parties and religious denominations in the one study. On the other hand, if preferences for social institutions of a similar type are being studied and if such institutions do excite strong positive and negative attitudes in the majority of people in the population studied, then the unfolding technique has considerable potentialities. It is unlikely that the number of such institutions will make the method impractical.

A second point of methodology, which does not seem to have been stressed, relates to the collection of data. The unfolding technique places great emphasis on the individual's specific ranking. For this reason care should be exercised to ensure that there is some consistency in each subject's ranking. It is recommended that the method of paired presentations be used and that only consistent (i.e., transitive) choices be used as the basis of a rank order used to establish the position of stimuli and subjects. Inconsistent records can usually be located in the space after it has been established. In a study of the four major political parties in Queensland only 6 inconsistent records were found in 185 subjects. The remaining 179 rank orders were accounted for by a (6, 6, 4, 2) solution with 4 discrepancies.

Appendix

The following outlines indicate how the various statements in the text may be proved.

Consider a point D inside the triangle ABC. Construct the perpendicular bisectors of DA, DB, DC meeting in E, F, and G (Fig. 6).

Any point inside the space defined by the angle EGF is closer to D than to A or C. Any point in the space defined by the vertically opposite angle is closer to D than to B. Similar statements may be made for angles GEF

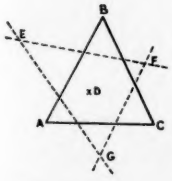


FIGURE 6

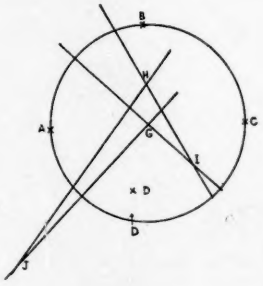


FIGURE 7

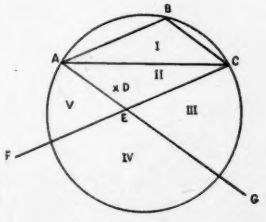


FIGURE 8

and EFG, and these three angles and their vertically opposite angles cover the total space. Thus the point D can never be further away from any point in the plane than all of the points A, B, C. This establishes the (6, 6, 6, 0) solution since there are 18 orders and no point can be placed last more than 6 ways.

Let D move to the line AC then the point G disappears as the bisectors of AD and DC become parallel. Thus a space in which B will be placed last disappears and as this space is divided by the perpendicular bisector of AC two orders ending in B will disappear to give the (6, 6, 4, 0) solution. If D drops below AC then G reappears above B to define two areas in which D may be placed last, thus establishing the (6, 6, 4, 2) solution.

Fig. 7 shows ABCD as cyclic. If D moves inside the circle defined by ABC then B will also be inside the circle ACD. If the perpendicular bisectors of AB, BC, CD, DA meet in G, H, I, and J, then any point inside GIHJ cannot be further from B or D than it is from both A and C, i.e., the six orders corresponding to subdivisions of GIHJ must end in A or C which can appear last in all six possible ways.

B and D have been taken as opposite points of the quadrilateral ABCD so that D can occupy only the regions I, II, III, IV, or V of Fig. 8 (or their boundaries which give degenerate solutions). Region I has been dealt with as the "concave" case; and it will be shown that regions II and IV and regions III and V are equivalent in pairs and correspond to (6, 6, 4, 2) and (6, 6, 3, 3) solutions, respectively. Thus the solution types (6, 6, 6, 0); (6, 6, 4, 2); (6, 6, 3, 3) are the only nondegenerate solutions.

Consider D outside the triangle ABC and complete the parallelogram ABCE. If D is inside ABCE no new perpendicular bisectors become parallel so that this is the (6, 6, 4, 2) solution. If D moves to one side of the parallelogram, say EC, then the perpendicular bisectors of AB and DC fail to meet and one order, ending in B disappears to give a (6, 6, 3, 2) solution. If D moves across this side into region III then these two bisectors meet above B and a further order ending in D appears to give a (6, 6, 3, 3) solution.

Clearly, if D is in region IV, B will be inside the circle ADC and the parallelogram ADC; and from similar configurations the sets of (6, 6, 4, 2) solutions arise.

REFERENCES

[1] Bennett, J. F. and Hays, W. L. Multidimensional unfolding: determining the dimensionality of ranked preference data. *Psychometrika*, 1960, **25**, 27-43.

Manuscript received 11/7/60

Revised manuscript received 2/27/61

PSYCHOMETRIKA—VOL. 26, NO. 3
SEPTEMBER, 1961

A NOTE ON A CLASS OF PROBABILITY MATCHING MODELS*

JULIAN FELDMAN
UNIVERSITY OF CALIFORNIA, BERKELEY

AND

ALLEN NEWELL
THE RAND CORPORATION

Probability matching is shown to be a property of a broad class of models of binary choice behavior.

In the *binary choice experiment*, the subject is asked to predict which of two possible events, E_0 or E_1, will occur on each of a series of trials. One event occurs on every trial; both events do not occur on the same trial. The sequence of events is usually determined by a random mechanism. The subject is generally not told anything about the method of constructing the series, but he is told which event did occur after he makes his prediction.

Estes and his associates [e.g., 7] and Bush and Mosteller ([4], p. 280) have offered models of binary choice behavior as special cases of their learning theories. These two models are mathematically identical although they stem from different psychological assumptions. The principal evidence offered for the validity of the Estes-Bush-Mosteller model has been the ability of the model to predict the phenomenon of *event matching* or *marginal probability matching*, i.e., the model predicts each event about as often as that event occurs in the event series; event matching has been observed in several experiments [e.g., 8, 9]. Edwards [5] has contended that subjects tend to predict the most frequent event more often than it actually occurs. Other investigators report that subjects predict the most frequent event more often than it occurs when correct predictions are rewarded with money and incorrect predictions are penalized with a loss of money [15] and when the events are playing cards drawn from a deck [13]. However these same investigators [13, 15] report event matching in their control conditions. Models that do not predict event matching have been offered by Luce ([11], ch. 4) and Siegel [14].

Simon [16] and Bryant and Marica [2] have also presented models of

*This note is based in part on a section of Feldman's doctoral dissertation submitted to the Graduate School of Industrial Administration, Carnegie Institute of Technology.

binary choice behavior derived from plausible assumptions. Both of these models are also consistent with event matching. In this note, a broad class of learning models, consistent with event matching and including the Estes-Bush-Mosteller model as a special case, will be presented.

The Estes-Bush-Mosteller model has been criticized because of its inability to predict *conditional probability matching*, which has been reported by some investigators [e.g., 1, 6, 9]; here the conditional probabilities of the subject's prediction series match the conditional probabilities of the event series. Burke and Estes [3] and Sternberg [17] have extended the original Estes-Bush-Mosteller model to enable it to predict conditional probability matching. Neither the Simon nor Bryant and Marica models can predict conditional probability matching. The behavior of the Bryant-Marica model depends on its ability to predict correctly and hence need not necessarily satisfy the conditional probability criterion. In the model proposed by Simon, the probability of predicting E_1 on trial t depends on the prediction on trial $t - 1$ and not on the event $t - 1$. A subclass of the class of models presented in this paper can predict conditional probability matching. This subclass includes the models of binary choice behavior offered by Hake and Hyman [9] and Nicks [12] as special cases when certain additional assumptions are made about these two models.

A General Class of Probability Matching Models

The class of models will be developed in two stages. First, consider the case where p_t , the probability that the subject will predict that E_1 will occur on trial t, is a moving average over the events preceding trial t and an initial condition, p_0 . The weights of the moving average, the w's, satisfy the conditions,

$$0 \leq w_i \leq 1, \qquad \sum_{i=1}^{\infty} w_i = 1.$$

Let $E_0 = 0$, $E_1 = 1$, and E_t be the event that occurs on trial t, $E_t = 1, 0$. Then

$$p_t = \sum_{i=1}^{t-1} w_i E_{t-i} + p_0 \sum_{i=t}^{\infty} w_i \,. \tag{1}$$

If $\Pr(E_1) = \pi$, then the expected value of p_t , $\bar{p}_t$, can be expressed in the following form:

$$\bar{p}_t = \pi \sum_{i=1}^{t-1} w_i + p_0 \sum_{i=t}^{\infty} w_i \,,$$

$$\bar{p}_t = \pi\left(1 - \sum_{i=t}^{\infty} w_i\right) + p_0 \sum_{i=t}^{\infty} w_i \,,$$

$$\bar{p}_t = \pi + (p_0 - \pi) \sum_{i=t}^{\infty} w_i \,.$$

If

$$\sum_{i=1}^{t-1} w_i \cong 1, \quad \text{then} \quad \bar{p}_t \cong \pi,$$

and the marginal probability requirement is satisfied for models represented by (1).

To obtain the complete class of models, consider a set of states, S_1 , S_2 , $\cdots$, S_m , such that each event can be uniquely assigned to a state. That is, at trial t, one and only one state S_k obtains. The state may be defined in any way consistent with unique assignment and the subject's ability to know the state, e.g., the state can be some function of past events or past predictions. The states partition the event series into m subseries. For each subseries, let E_j^k be the event following the jth occurrence of state S_k . Let π^k be the conditional probability of the event E_1 given the state S_k . The general model is formed by allowing the subject to have a separate moving average determine behavior after the occurrence of each state. If p_j^k is the conditional probability of predicting E_1 after the jth occurrence of state S_k , then for each subseries

$$p_j^k = \sum_{i=1}^{j-1} w_i^k E_{j-i}^k + p_0^k \sum_{i=j}^{\infty} w_i^k . \tag{2}$$

Equation (2) is of the same form as (1); thus, if

$$\sum_{i=1}^{j-1} w_i^k \cong 1, \quad \text{then} \quad \bar{p}_j^k \cong \pi^k,$$

and the conditional probability matching criterion is satisfied independently of the form of the distribution of the w's.

The marginal probability of event E_1 , π, can be expressed as

$$\pi = \sum_{k=1}^{m} \pi^k \Pr(S_k).$$

The marginal probability of predicting E_1 , $\bar{p}_j$, can be expressed as

$$\bar{p}_j = \sum_{k=1}^{m} \bar{p}_j^k \Pr(S_k).$$

Since $\bar{p}_j^k \cong \pi^k$, then $\bar{p}_j \cong \pi$, and the marginal probability criterion is satisfied.

Whether or not these models would reflect, independently of the w's, all of the conditional probability structure present in the event series would depend on the selection of the states. Consider the following example. (i) The event series consists of a double alternation, $E_1E_1E_0E_0E_1E_1E_0E_0 \cdots$. (ii) S_1 is E_1 and S_2 is E_0 . (iii) The w's take the form of an exponentially weighted moving average. The model accurately predicts the first-order conditional probabilities, i.e., Pr ($E_1 \mid E_1$), Pr ($E_1 \mid E_0$), Pr ($E_0 \mid E_1$), and Pr ($E_0 \mid E_0$).

However the model would not reflect the second-order conditional probabilities of the event series. If the states were E_0E_0 , E_1E_0 , E_0E_1 , and E_1E_1 , the model would reflect the second-order conditional probabilities [cf. 10].

Relation to Other Models

The general class of models presented above includes several of the previously proposed models of binary choice behavior. The original Estes-Bush-Mosteller model for binary choice behavior ([4], p. 280; [7]) is a special case of (1), where p_t is an exponentially weighted moving average over the events preceding trial t.* In the notation used here, the Estes-Bush-Mosteller model becomes

$$p_t = \alpha p_{t-1} + (1 - \alpha)E_{t-1} , \qquad 0 \leq \alpha < 1. \tag{3}$$

The general solution to (3) is

$$p_t = \alpha^{t-1}p_0 + (1 - \alpha) \sum_{i=1}^{t-1} \alpha^{i-1}E_{t-i} . \tag{4}$$

But (4) is just (1) with

$$w_i = (1 - \alpha)\alpha^{i-1} \quad \text{and} \quad \sum_{i=t}^{\infty} w_i = \alpha^{t-1}.$$

The models suggested by Hake and Hyman [9] and Nicks [12] can be considered statements of the states required by equation (2). Hake and Hyman suggest that the state for trial t consists of the two preceding events and the two preceding predictions. Nicks suggests that the state for trial t consists of the run of like events preceding trial t. Neither Hake and Hyman nor Nicks specify exactly how the prediction is obtained after the occurrence of each of these states. If the predictions are obtained in a manner consistent with (2), the Hake-Hyman model and the Nicks model will be consistent with marginal probability matching and conditional probability matching.

Summary

Recent work on binary choice behavior has resulted in the development of several models that are consistent with the evidence on event matching. The evidence on conditional probability matching has caused modification of some models and suggests rejection or modification of other models. The class of models in this note is consistent with event matching, and a subclass is consistent with conditional probability matching and includes several other models of binary choice behavior as special cases. Nevertheless, we suggest that additional tests be devised and additional empirical work be

*We are indebted to C. Holt for this observation.

done to specify the structure and parameters of a model or models of binary choice behavior.

REFERENCES

[1] Anderson, N. H. and Grant, D. A. A test of a statistical learning theory model for two-choice behavior with double stimulus events. *J. exp. Psychol.*, 1957, **54**, 305–317.

[2] Bryant, S. J. and Marica, J. G. Strategies and learning models. *Psychometrika*, 1959, **24**, 253–256.

[3] Burke, C. J. and Estes, W. K. A component model for stimulus variables in discrimination learning. *Psychometrika*, 1957, **22**, 133–145.

[4] Bush, R. R. and Mosteller, F. *Stochastic models for learning*. New York: Wiley, 1955.

[5] Edwards, W. Reward probability, amount, and information as determiners of sequential two-alternative decisions. *J. exp. Psychol.*, 1956, **52**, 177–188.

[6] Engler, J. Marginal and conditional stimulus and response probabilities in verbal conditioning. *J. exp. Psychol.*, 1958, **55**, 303–317.

[7] Estes, W. K. and Straughan, J. H. Analysis of a verbal conditioning situation in terms of statistical learning theory. *J. exp. Psychol.*, 1954, **47**, 225–234.

[8] Grant, D. A., Hake, H. W., and Hornseth, J. P. Acquisition and extinction of verbal conditioned responses with differing percentages of reinforcement. *J. exp. Psychol.*, 1951, **42**, 1–5.

[9] Hake, H. W. and Hyman, R. Perceptions of the statistical structure of a random series of binary symbols. *J. exp. Psychol.*, 1953, **45**, 64–74.

[10] Kochen, M. and Galanter, E. H. The acquisition and utilization of information in problem solving and thinking. *Information and Control*, 1958, **1**, 267–288.

[11] Luce, R. D. *Individual choice behavior: a theoretical analysis*. New York: Wiley, 1959.

[12] Nicks, D. C. Prediction of sequential two-choice decisions from event runs. *J. exp. Psychol.*, 1959, **57**, 105–114.

[13] Rubinstein, I. Some factors in probability matching. *J. exp. Psychol.*, 1959, **57**, 413–416.

[14] Siegel, S. Theoretical models of choice and strategy behavior: stable state behavior in the two-choice uncertain outcome situation. *Psychometrika*, 1959, **24**, 303–316.

[15] Siegel, S. and Goldstein, D. A. Decision-making behavior in a two-choice uncertain outcome situation. *J. exp. Psychol.*, 1959, **57**, 37–42.

[16] Simon, H. A. A comparison of game theory and learning theory. *Psychometrika*, 1956, **21**, 267–272. (Reprinted in H. A. Simon, *Models of man*. New York: Wiley, 1957. Ch. 16.)

[17] Sternberg, S. H. A path-dependent linear model. In R. R. Bush and W. K. Estes (Eds.), *Studies in mathematical learning theory*. Stanford: Stanford Univ. Press, 1959. Ch. 16.

Manuscript received 5/20/60

Revised manuscript received 11/7/60

BOOK REVIEWS

MORDECAI EZEKIEL AND KARL A. FOX. *Methods of Correlation and Regression Analysis.* Third Edition. New York: John Wiley and Sons, 1959. Pp. xi + 548.

Some future historian, concerned with the statistical fads and fashions of our era, may well decide that the third edition of Ezekiel's influential book should have been titled "The rise and fall of the correlation coefficient." The first edition, in 1930, helped to accelerate the use of the Pearson product-moment coefficient. However, even in the first edition, and more so in the 1941 second edition, there was a pragmatic emphasis on regression coefficients, errors of prediction, nonlinear relations, and graphic curve-fitting, areas where the exact value of the correlation coefficient is not too important. In the preface to the present edition, Ezekiel and Fox state that their "major emphasis . . . has shifted from correlation to regression." The book jacket blurb mentions correlation only once, and then in connection with an effort to minimize its importance. "This book contains a comprehensive treatment of all aspects of regression analysis, and of correlation analysis insofar as the latter contributes to the description and evaluation of regression studies."

Accordingly, it is no surprise to find that the major new additions by Fox are (i) a chapter on fitting systems of simultaneous equations, (ii) a chapter covering "only certain aspects of variance analysis that are closely related to regression problems," (iii) changes to make the standard error formulas for regression coefficients conform more closely to current terminology, and (iv) autocorrelation tests and von Neumann ratio tests for regression residuals from time series.

What has happened to Ezekiel's unique (and controversial) sections on freehand fitting of curves for multivariate nonlinear regressions? They are still there, but an attempt to estimate the number of constants has been made, and a hint is given to the astute reader that the successive graphic approximations do not necessarily converge. It is suggested that electronic computers can work so fast that it is possible to explore a wide variety of algebraic equations, thus reducing the usefulness of graphic successive approximations, although this is followed by an explicit statement of how to use electronic computers to speed up the graphic approximations. However, compared to the second edition, there has been a definite shift away from freehand fitting methods. To quote the jacket blurb again, "primary emphasis is placed upon algebraic methods of determining regression. Graphic methods . . . are also fully treated."

Another unique feature that has been retained in part is the discussion of "joint regression," where the regression of the criterion on one or more independent variables is itself a function of one or more of the independent variables. Algebraically, this leads to the same sort of nonlinear equation that results from the influence of a "moderator" variable like sex or age on a criterion-predictor regression. Unfortunately, the description of Court's generalization has not been retained (Andrew T. Court. Measuring joint causation. *J. Amer. statist. Ass.*, 1930, **25**, 245–254). In general, Court fits the criterion to a quadratic function of the independent variables. If the criterion can be adequately predicted by a linear multiple regression in which the regression coefficients are *linear* functions of the independent variables, then a general quadratic equation is sufficient to describe the criterion-predictor relation. If the coefficients of the linear multiple regression are *nonlinear* functions of the independent variables, the general equation will be a multivariate power series, i.e., a multivariate Taylor series expansion. Configural analysis (H. G. Osburn and A. Lubin. A theory of pattern analysis for the prediction of a quantitative criterion.

Psychometrika, 1957, **22**, 63–73) can be shown to be a special case of the multivariate Taylor series expansion.

All the material on the characteristics of parabolic, logarithmic, and hyperbolic functions has been retained. Some reviewers of the first and second editions felt that the discussion was too brief, contained some mathematical errors, and omitted important equations such as the exponential function. Although the exponential function is not discussed, its logarithmic form is. The fifty pages or so which discuss how, when, and where curvilinear functions should be used may indeed be too brief for a book devoted to regression analysis, but where is the text which devotes more space to the logic and practice of curve fitting?

After emphasizing the importance of logical bases for the selection of particular mathematical functions in curve fitting, Ezekial and Fox argue that where no such logical bases can be developed "a curve fitted freehand by graphic methods, and conforming to logical limitations on its shape, may be even more valuable as a description ... than a definite equation ..." which fits less well. This overlooks several important virtues of graphic methods. If all that is wanted is a good fit, most research workers have access to computers that can mechanically generate power series which will give any desired degree of goodness of fit. Surely the main advantages of the graphic methods are that the research worker looks at the data and can decide whether aberrant observations are present which conflict with group trends, can see whether some a priori equation is clearly not appropriate, can spot scoring errors, etc. Graphic methods make it possible to grasp all or most of the data simultaneously. Not only is it possible to exclude large numbers of postulated relations but often certain functions and possible causative mechanisms seem to be suggested by the graphs.

On the whole, the third edition is not much better than the second edition. Tests of significance for multiple correlations and regression coefficients have been corrected. The additional sections on systems of simultaneous equations and on time series tests are interesting, but these subjects have been treated more extensively in other texts. The interpolated sections on analysis of variance (which is barely mentioned in the second edition) are only superficially related to the rest of the book. It would be possible (and desirable) to derive all methods of regression analysis from the concept of the general linear hypothesis, i.e., to show that the form of regression analysis depends on whether the variables are quantitative, qualitative, or a mixture. However, this would demand a fundamental rewriting. There is a certain patchiness and redundancy in the book at present which could be eliminated if it were organized about the general linear hypothesis.

Because there is no common network of concepts, each model is presented *de novo*, divorced from other models and from real situations where the stated assumptions will not be met. This implies artificial restrictions on the use of some models. For example, in the correlation model described by Ezekiel and Fox, the joint distribution of X and Y is bivariate normal. They do not mention that for many non-normal joint distributions (e.g., rectangular), as N increases, the sampling distribution of correlation coefficients will asymptotically approach that of the bivariate normal distribution.

I am indebted to Prof. John H. Smith, American University, for pointing out that the Ezekiel and Fox regression model is more general than it sounds. Their regression model has no restrictions on the distribution of X, but successive samples are assumed to have the same values of X, with values of Y being normally distributed with each X array. Smith points out that "the deviation of the sample regression coefficient from the hypothetical universe value is 'standardized' to obtain the criterion for the t-test in exactly the same way as for the criterion coefficient. For the regression model it is possible to state correctly that the denominator of this criterion is the square root of an independent unbiased estimate of the sampling variance of the regression coefficient. A similar statement is not true for the correlation model. In this sense, the exposition for the regression model is

simpler even when techniques turn out to be identical. This is true also for interval estimation of regression coefficients.

"The regression model is inadequate for dealing with inferences as to values of correlation coefficients in normal bivariate universes when these coefficients are not equal to zero. However, the criterion for making the t-test of the hypothesis that $\rho = 0$ is the same for the regression model as for the correlation model. In this case, the t-function of r is not equal to the ratio of r to an estimate of its standard error. The exact sampling variance of r for uncorrelated universes is known to be $1/(n - 1)$, where n is the number in the sample. The variance of r and, in fact, the whole sampling distribution of r, is the same for regression and correlation models provided only that the universe correlation is zero, regression is linear, and arrays are normal and homoscedastic."

I would not recommend the Ezekiel and Fox book as a text in any class, but the research worker who uses regression analysis would be wise to keep a copy handy for reference to the unique sections on joint functions and graphic multivariate curvilinear analysis.

Walter Reed Army Institute of Research

ARDIE LUBIN

ALEXANDER S. LEVENS. *Nomography*. Second Edition. New York: John Wiley and Sons, 1959. Pp. viii + 296.

The author has set himself the task of outlining the basic theory and method of construction of various types of nomographs involving straight-line as well as curved scales. The first two chapters of the book are devoted to an introduction to the general theory and method as well as a discussion of methods for converting numerical values to points on related geometric scales. Chapters three through eleven then describe the various geometric forms appropriate for use in solving a wide variety of types of equations. Chapters twelve through sixteen represent substantial additions to the author's first edition and include a chapter each on the design of net charts, circular nomograms, and an expanded discussion of the use of determinants in the construction of alignment charts. There is also a chapter on projective transformations, and another on the relationship between concurrency (Cartesian) and alignment nomographs.

Any psychologist who is interested in learning about the construction of nomographs will find this book extremely useful both from the point of view of a general discussion of theory, and as a reference source for developing a set of scales for solving any particular formula which must be repetitively applied.

The only reservation which this reviewer has in evaluating the book from the viewpoint of the psychologist has to do with the appendix, which includes 58 different nomographs showing examples of the use of nomography in a wide variety of different fields. Six of these 58 have direct applications for psychological statistics. Unfortunately, the author's choice is such that the casual reader might easily be left with the impression that nomographs have only limited practical use in psychological statistics. Historically, nomographs in psychology have found their most valuable use in the rapid and repetitive calculation of various types of fourfold correlations, or in testing the significance of the difference between groups using dichotomous data. In sharp contrast, the author presents as a useful example, a nomograph for calculating a product moment correlation using $\sum x^2$, $\sum y^2$, and $\sum xy$ in deviation units. Anyone who would take the trouble to calculate these values would question whether use of the nomograph represents a saving of time over the more conventional computational formulas. Another example is a nomograph for the Spearman-

Brown formula for estimating the reliability of a lengthened test. In the reviewer's opinion this also is not a very useful example since the Spearman-Brown formula is rarely an equation which has to be repetitively solved. The reviewer's concern is that the uninitiated psychologist, being confronted with nomography for the first time, may not be immediately aware of the practical and time-saving advantages which can accrue from a knowledge of the theory and application of the method, which this book so adequately describes.

Marketing, Merchandising and Research, Inc. VALENTINE APPEL
New York City